New forensic DNA profiling techniques for human identification

Felicia Bardan

Australian Centre for Ancient DNA
School of Biological Sciences
Faculty of Sciences
University of Adelaide

A thesis submitted in fulfillment of the requirements for the degree of Doctor of Philosophy at the University of Adelaide

February 2019

Table of Contents

Thesis Abstract.....i
Thesis Declaration..... iii
Acknowledgements iv

Chapter 1: General Introduction 1

1.1 Conventional Methods for Human Identification.....2
1.1.1 Non-DNA Based Identification 2
1.1.2 Conventional DNA-based Identification: Short Tandem Repeats.....3

1.2 The Influence of DNA Quality and Quantity on DNA Typing Success.....4

1.3 Previous Advancements in Forensic DNA Typing.....6
1.3.1 MtDNA Control Region for Human Identification 6
1.3.2 Single Nucleotide Polymorphisms 7

1.4 Limitations of mtDNA and SNPs for Forensic Investigation.....11
1.4.1 Multiplex Assay Design 11
1.4.2 Lack of Representative Population Reference Databases 13

1.5 Emerging Technologies for Degraded DNA Analysis 15
1.5.1 Current Commercial MPS Technologies for Human Identification.....16
1.5.2 Hybridisation Enrichment as an Alternative Approach for Typing Degraded DNA 17

1.6 Overview of Thesis and Data Chapter Summaries.....18

References.....21

Chapter 2: A mini-multiplex SNaPshot assay for the triage of degraded human DNA..33

Chapter 3: A custom hybridisation enrichment forensic intelligence panel to infer biogeographic ancestry, hair and eye colour and Y-chromosome lineage 53

Abstract.....56

Introduction.....57

Materials and Methods.....58

Results.....71

Discussion 82
Conclusion 87
References 88
Supplementary Information 94

Chapter 4: Application of the Miniplex SNaPshot assay and the 124-SNP hybridisation enrichment assay to degraded human DNA 109

Abstract 112
Introduction 113
Materials and Methods 114
Results 118
Discussion 141
Conclusion 146
References 147
Supplementary Files 151

Chapter 5: The Historical Australian DNA Database 159

Abstract 162
Introduction 163
Materials and Methods 167
Results 171
Discussion 177
Conclusion 182
References 183
Supplementary Information 189

Chapter 6: General Discussion and Conclusion 203

Introduction and Thesis Summary 204
Significance 206
Broader Applications 211

Limitations and Recommendations for Future Directions 214

Concluding Remarks 223

References 224

Thesis Abstract

Highly degraded biological samples are commonly encountered in missing persons cases, historical human remains, war graves, mass disasters and various forensic casework. As biological tissue degrades, DNA becomes progressively fragmented and chemical modifications can occur, complicating successful standard short tandem repeat typing. Alternative genotyping strategies such as single nucleotide polymorphism typing and the emergence of massively parallel sequencing to examine ancestry and phenotype SNPs have ushered in a new era of forensic intelligence testing for problematic samples. Despite showing promise, a number of technical concerns still exist for the use of these strategies in forensic investigation.

The research presented in this thesis explores, develops and assesses alternative techniques using both traditional and new technologies for the retrieval of forensic intelligence data from highly degraded samples. I develop new techniques for the screening and genotyping of highly degraded DNA and generate a new dataset of ancestry data from an Australian population for use in analysing historical samples. Issues relating to the implementation of these technologies are discussed, including laboratory workflow, data analysis and interpretation, ethics, and the need for standard guidelines for forensic laboratories to adopt in their methodology.

Specifically, in this thesis I use:

- A SNP typing strategy based on conventional techniques and equipment to develop a screening tool that estimates sample degradation and presumptive broad biological profile for the triage of forensic samples – Chapter 2
- Emerging target enrichment and massively parallel sequencing technologies for the generation of ancestry and phenotype data for forensic investigation – Chapter 3
- Techniques developed and assessed in Chapter 2 and 3 to analyse a set of degraded DNA and forensic casework samples, demonstrating the utility of the methods to genotype and provide forensic intelligence data for challenging samples – Chapter 4
- mtDNA and autosomal SNP analysis to construct the first Australian reference population database for ancestry testing of historical human remains – Chapter 5

In essence, my research aimed to explore techniques to improve the genetic assessment of highly degraded and compromised forensic samples, and to build on current knowledge concerning the implementation of such techniques in forensic investigations.

Thesis Declaration

I, Felicia Bardan, certify that this work contains no material which has been accepted for the award of any other degree or diploma in my name, in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text. In addition, I certify that no part of this work will, in the future, be used in a submission in my name, for any other degree or diploma in any university or other tertiary institution without the prior approval of the University of Adelaide and where applicable, any partner institution responsible for the joint-award of this degree.

I acknowledge that copyright of published works contained within this thesis resides with the copyright holder(s) of those works.

I also give permission for the digital version of my thesis to be made available on the web, via the University's digital research repository, the Library Search and also through web search engines, unless permission has been granted by the University to restrict access for a period of time.

I acknowledge the support I have received for my research through the provision of an Australian Government Research Training Program Scholarship.

26/02/2019

..... ….....

Felicia Bardan Date

Acknowledgements

I've always wanted to be a forensic scientist. Since I was just 14, I knew what I wanted and worked for years towards it. But it had never occurred to me that I'd have the opportunity or tenacity to complete a PhD. Yet I committed myself to the challenge when presented and here I am, but it's not without the tremendous support from my family, friends and colleagues, and mostly my supervisors. To all of you, thank you for believing in me and sticking it out, even in times when that belief would have been questionable. The guidance and professional leadership from my supervisors and co-authors Jeremy Austin and Denice Higgins offered was invaluable for my development. Thank you for giving me a chance. I would also like to extend my thanks to the University of Adelaide for funding my candidature and for the Australian Research Council for the financial support for the projects to commence. For all my colleagues and friends from ACAD, my PhD journey may have been far darker if it weren't for the consistent support both emotionally and professionally. Leanne van Weert, Jennifer Young, Lauren White, Emily Skelly, and Caitlin Selway, cheers to you for your constant willingness to have a laugh (or cry) over a G&T. To my loyal furry and feathery companions Vulcan and Honey, I am indebted to you for the endless hours of companionship and comfort sitting by my feet and perching on my shoulder. And finally, my family and remarkable partner Sam, I cannot thank you enough for being so selfless in your sacrifices during my years of study. Your unwavering love, faith and encouragement will never be forgotten or left unappreciated, I promise. I also want to specifically honour my parents and grandparents, who risked their lives to seek liberty for themselves and their future families. You've made it possible for me to live a happy and healthy life, and ultimately, allowed me to chase my dreams.

> *"The stories the surviving defectors tell are chilling. The grit some showed may seem borderline to insanity... Their stories are important, no matter if they made it to the West or died trying… They dared to hope."*
>
> - Marina Constantinou

My life would not be as rich and full of love as it is, and my PhD would not have been completed as it was without you all.

Chapter 1

General Introduction

Forensic human identification is the individualisation of human remains by attribution of a legal name and is required for confirming the identity of victims of crime and armed conflicts, missing persons, disaster victims, and for the settlement of estates. Identification of human remains is therefore important for both legal and social reasons. Incidents such as homicide (437,000 people globally in 2014), missing persons (38,000 Australians annually with 2000 long term - missing for >3 months), and natural disasters (e.g. 230,000 victims of the Boxing Day tsunami) all result in a large number of human remains (Akkoc 2014; United Nations Office of Drugs and Crime 2014; Bricknell 2017; Ward 2018). In such circumstances, forensic practitioners are tasked with the responsibility of assigning a legal identity to the remains.

1.1 Conventional Methods for Human Identification

There are a number of conventional approaches (both physical and genetic) that can be used to identify an individual. The use of physical evidence for identification (e.g. fingerprints, dentition) is not always successful or even plausible for remains that are highly fragmentary or decomposed. In some cases, identifying physical features are no longer intact due to advanced decomposition or trauma resulting from incineration or high impact events. For these situations, DNA evidence can be the only remaining source of establishing human identity.

1.1.1 Non-DNA Based Identification

Conventional human identification methods not based on genetic analysis include visual identification by someone known to the deceased, as well as the comparison of physical features such as fingerprints, medical implants, dentition and tattoos to ante-mortem records. These forms of human identification show high success on human remains that possess the characteristics required for analysis (for example, a skull with retained teeth for dental comparison) (Wright et al. 2015).

An example of an international relief effort where non-DNA based evidence was successful can be seen in the 2004 tsunami disaster off the coast of Thailand where over 5000 people from 44 countries were killed (Akkoc 2014). Hundreds of international forensic odontologists, pathologists, biologists, anthropologists and fingerprint examiners participated in a scientific and humane endeavour to determine the identity of the victims through the use of primary forensic evidence (odontology, fingerprints, physical features and DNA). Odontology and fingerprint evidence were initially the preferred methods for identifying victims due to the

costs, logistics and time constraints of DNA analysis. By 2008, odontology had contributed to 40% of positive identifications by comparing post-mortem dental evidence to ante-mortem dental records (Wright *et al.* 2015). Fingerprint evidence accounted for 35%, whereas DNA and physical evidence resulted in 24 and one percent of positive identifications respectively (Wright et al. 2015). DNA-based identification resulted in a low number of positive identifications during the early to midstages of the repatriation process, however grew in success in the later stages for remains which proved to be problematic for analysing dental, fingerprint and physical evidence (Wright et al. 2015). This demonstrates that for such situations, DNA-based identification becomes a more viable option when other primary identification methods are exhausted or not possible due to a lack of diagnostic features.

1.1.2 Conventional DNA-based Identification: Short Tandem Repeats

The majority of the nuclear genome is almost identical in all people, with only 0.3% of its sequence differing between individuals (Butler 2005). The key aspect of DNA profiling for forensic identification is to target and analyse the genetic variations that exist from person to person. Forensic biologists primarily rely on the typing of Short Tandem Repeat (STR) regions owing to their high mutation rates and variability between individuals (Brinkmann *et al.* 1998; Amorim & Pereira 2005). The combination of these multi-allelic loci (multiplexing) results in a highly discriminatory technique that can be used to differentiate individuals from each other. STRs targeted in forensic identification are either autosomal STRs (Mattayat *et al.* 2016) for individualisation, on the Y chromosome (Y-STRs) that can infer biological sex and characterise the paternal lineage of a male donor (Gopinath *et al.* 2016), or - although less routine - are X-linked STRs that are useful for kinship testing (Barbaro *et al.* 2006; Israr *et al.* 2014).

STR typing via polymerase chain reaction (PCR) is currently the preferred method for DNA-based human identification (Butler 2011; Zietkiewicz *et al.* 2012). STR profiles generated from samples of unknown origin are compared directly against a curated reference database that includes STR profiles from convicted criminals, crime scene samples, persons of interests, or family members of a deceased/missing individual. The significance of a match is then given a level of statistical power by assessing the profile within a population database that estimates allele frequencies in a representative population (Taylor *et al.* 2017). The development of commercial STR kits (Wang *et al.* 2012; Applied Biosystems 2016) has resulted in global

standardisation and validation of these technologies and as such, the majority of forensic DNA databases worldwide are STR based (Alonso *et al.* 2005).

1.2 The Influence of DNA Quality and Quantity on DNA Typing Success

The DNA available for biological profiling of forensic samples is not always of sufficient quality and quantity for successful analysis due to degradation and the nature of the DNA molecules. Biological decomposition of tissues is influenced by two primary factors: environmental conditions and post-mortem interval (Burger *et al.* 1999). The manner in which an individual died can also influence the survival of tissues and the DNA within them (Schwark *et al.* 2011). PCR is currently the method of choice for in vitro amplification of DNA molecules but can fail to produce usable profiles when the DNA sample is not of adequate quality and quantity (Alaeddini *et al.* 2010).

Frequently, forensic samples have been exposed to harsh environmental conditions which affects DNA preservation and can lead to the accumulation of inhibitory substances, impacting genetic profiling. DNA repair mechanisms that maintain genome integrity in living cells no longer function after cell death (Lindahl 1993; Dabney *et al.* 2013). The degree of DNA degradation depends largely on the biological source and deposition environment and accumulates over time. Factors such as temperature, pH and humidity influence the rate and intensity of degradative processes (Fondevila *et al.* 2008a; Dabney *et al.* 2013). Warm and wet conditions promote microbial infestation and reactive oxygen species that alter DNA bases (Levy-Booth *et al.* 2007; Alaeddini *et al.* 2010). The chemical processes which fragment DNA sequences and breakdown the sugar-phosphate backbone are also accelerated by heat and humidity (Dabney *et al.* 2013). Increased UV radiation in regions of intense heat and sun exposure can also damage the DNA in remains that are unburied and unprotected from the elements. Photochemical exposure can induce the formation of covalent linkages between adjacent C or T bases along the DNA strand, resulting in pyrimidine dimers that can cause DNA polymerases to stall during PCR replication (Goodsell 2001). Immersion in water, fire, or burial in soil are also elements which can affect biological decomposition (Crainic *et al.* 2002; Higgins *et al.* 2015; Bogas *et al.* 2016). The presence of PCR inhibitors from the environment (e.g. humic substances from soil) can accumulate during decomposition and will also interfere with DNA typing success (Tsai & Olson 1992; Matheson *et al.* 2010).

Rather than a single insult, the composition and degradation of biological tissues and the DNA within them is a multi-factorial process resulting in challenging samples for genetic profiling. Whilst the mechanisms and degree of DNA damage and degradation stochastically differ between samples, the effect is the same: fragmentation of DNA sequences into shorter and shorter segments, and compromised DNA structure. This makes amplification of target sequences difficult, and often results in partial profiles or complete failure of genetic profiling. Different methods of DNA identification have different degrees of tolerance for sample quality and DNA quantity, and careful decisions must be made to better manage resources when analysing degraded material in order to maximise recovery of DNA whilst avoiding loss of sample.

While STR typing of nuclear DNA is well established in forensic identification, it provides little investigative value when there are no matches to a reference sample or database. Secondly, it has limited success with highly degraded remains (Mulero *et al.* 2008; Bogas *et al.* 2015). Typically, amplification of STR loci targets DNA sequences between 100bp-500bp (Butler 2007), and requires at least ~80 intact cells to obtain the optimal amount of DNA for successful typing (Kline *et al.* 2005). The process of post-mortem DNA damage causes DNA to fragment into sequences typically shorter than the amplicons in STR typing kits, and thus can interfere with PCR amplification success. Chemical modifications to the molecular structure of DNA also results in blocking lesions that cannot be bypassed by DNA polymerases, leading to amplification failure (Lindahl 1993; Goodsell 2001; Sikorsky *et al.* 2007; Nelson 2009; Shafirovich & Geacintov 2010; Dabney *et al.* 2013). Analysis of degraded DNA often produces partial profiles where the larger loci fail to amplify, termed the 'ski-slope effect' (Opel *et al.* 2006), or in extreme cases, can completely fail depending on the extent of DNA damage and thus impede investigations of identity. The endogenous DNA (authentic DNA from the individual) in forensic samples can vary from high to low quality and quantity. These variations are important factors in the success of DNA identification and need to be considered to determine which techniques are likely to be successful. For samples that have suffered prolonged exposure to sub-optimal environments where DNA damage processes are highly active, other DNA-based identification techniques are continuing to be explored and applied.

1.3 Previous Advancements in Forensic DNA Typing

Ongoing developments in molecular biology techniques are providing new avenues to retrieve genetic material and successfully genotype challenging forensic samples. These have focused mostly on the analysis of alternative genetic marker types to those used in conventional STR typing to aid in human identification attempts (Edson *et al.* 2004; Musgrave-Brown *et al.* 2007; Fondevila *et al.* 2008b; Coble *et al.* 2009).

1.3.1 MtDNA Control Region for Human Identification

Forensic scientists rely on mitochondrial DNA (mtDNA) to generate biological profiles from compromised samples that fail with standard STR typing due to insufficient template quality and quantity (Holland *et al.* 1993; Budowle *et al.* 2003). In contrast to nuclear DNA, a single human cell contains multiple copies of the mitochondrial genome. mtDNA is inherited maternally without recombination, so maternal relatives will share the same mtDNA lineage, making it useful for kinship testing. Thus, a maternal relative can be used as a reference sample for identification, or for exclusionary purposes (Hartman *et al.* 2015). MtDNA is more likely to be preserved in degraded tissues because of its robust circular structure, which is thought to impart some resistance to DNA degradation that nuclear DNA lacks (Budowle *et al.* 1999; Butler 2011). Its existence in higher copy number per cell relative to nuclear DNA also means it is more likely to be recovered from degraded biological material (Legros *et al.* 2004; Butler 2005; Foran 2006). For these reasons, mtDNA is usually investigated for remains that are degraded and possess low quantities of DNA.

MtDNA variation between individuals is commonly analysed by sequencing two hypervariable segments (HV1 and HV2) of the control region (Holland *et al.* 1993; Budowle *et al.* 2003; Edson *et al.* 2004; Parson & Dur 2007). Because the control region is non-coding, nucleotide variability is more abundant (Butler 2009). While mtDNA has become a valuable tool for human identification, mtDNA is a haploid genome providing less power of discrimination than STRs, but is useful for testing familial relationships (Coble *et al.* 2009). The forensic European DNA Profiling Group (EDNAP) mtDNA Population Database (EMPOP) (Parson & Dur 2007) estimates that approximately 7% of Europeans share the most common control region haplotype (Allard *et al.* 2002; Coble *et al.* 2004). Variations in the mtDNA coding region can be examined to increase the resolution, discriminatory power and confidence of a haplotype match or exclusion (Parsons & Coble 2001; Quintans *et al.* 2004; Fridman *et al.* 2011).

Currently however, coding region polymorphisms are not routinely examined and reported for forensic human identification since they are not widely implemented.

1.3.2 Single Nucleotide Polymorphisms

Single nucleotide polymorphism (SNP) typing offers an alternative method for investigations in cases involving unsuccessful STR typing. SNPs can be valuable in genotyping highly degraded DNA due to the ease of reducing amplicon sizes (Budowle 2004; Phillips *et al.* 2004). SNPs constitute approximately 90% of genomic variation and exist across the whole human genome (Collins *et al.* 1998). SNPs can be interrogated for a range of different purposes for forensic investigation, including individualisation (Musgrave-Brown *et al.* 2007), and gathering of intelligence information via the prediction of biogeographic ancestry and pigmentation of hair, eyes and skin (de la Puente *et al.* 2016; Chaitanya *et al.* 2018).

1.3.2.1 Identity Informative SNPs

Identity informative SNPs can be targeted as an alternative means for individual identification when testing challenging remains (Freire-Aradas *et al.* 2012). SNPs for identification are required to have high heterozygosity and low population heterogeneity (Budowle & van Daal 2008). The vast majority of SNPs are bi-allelic. Because of this, they are much less discriminatory than multi-allelic STRs on a per-locus basis, and studies suggest that between 50 to 80 identity SNP markers would need to be analysed to match the discrimination power of a 10-16 STR locus panel (Gill 2001).

The SNPforID consortium has constructed a 52-plex SNP assay to aid in human identification, where maximum amplicon size is 115bp, making it more suitable for use on degraded DNA (Sanchez *et al.* 2006). Other researchers have also constructed identification SNP panels where the selected autosomal markers collectively give very low probabilities of two individuals having the same multi-locus genotype (Dixon *et al.* 2005; Pakstis *et al.* 2010; Butler 2011).

1.3.2.2 SNPs for Intelligence Gathering

SNPs interrogated for ‘intelligence’ do not result in direct confirmation of identity but can provide information to lead investigators to a targeted search for positive identification. This is beneficial when individual identification cannot be established from STR profiling (either from no matches to a database or where no profile can be obtained from degraded remains). Selected SNP markers that exist on nuclear and mitochondrial DNA have been used in forensic

investigation of genetic ancestry because some alleles are associated with specific populations (Phillips *et al.* 2007; Butler 2011; Kayser & de Knijff 2011; van Oven *et al.* 2011b; Valverde *et al.* 2013), or are associated with differences in hair and eye colour (Sulem *et al.* 2008; Walsh *et al.* 2013). The predictive capability of DNA can act as a 'genetic witness' to provide complementary forensic intelligence for cold cases and missing persons, and for cases where investigative leads have been expended (Phillips *et al.* 2009; Chaitanya *et al.* 2017).

1.3.2.2.1 Phenotype-informative SNPs

Genome-wide association studies have revealed genes involved in complex traits such as external visible characteristics (EVCs) (Han *et al.* 2008; Sulem *et al.* 2008). Eye, hair and skin colour are highly heritable traits and are important for human identification as they are key visual descriptors of an individual. The ability to determine visual characteristics from DNA can be a valuable tool in helping to solve missing person cases, identifying mass disaster victims or historical skeletal remains by providing intelligence information when other physical and genetic information is limited (Chaitanya *et al.* 2017). Human pigmentation in hair and eye colour is a polygenic complex trait, determined by the combined effects of multiple genes that control melanin synthesis or localisation (Sturm *et al.* 2001). Numerous SNPs have been identified by previous studies to be strongly associated with differences in hair and eye colour (Shekar *et al.* 2008; Sturm *et al.* 2008; Sulem *et al.* 2008). The HIrisPlex system (Walsh *et al.* 2013) was developed as a single multiplex assay based on SNaPshot™ chemistry targeting 24 loci that distinguish hair colour (blond, brown, red and black) and eye colour (brown, intermediate and blue) with reasonable accuracy (>69.5% for hair colour, >82% for eye colour). This system uses a statistical prediction model to classify an individual into a phenotype class and provide an associated probability value by comparison to a reference database. A more recent panel, the HIrisPlex-S system types an additional 17 markers that are predictive for skin colour (Chaitanya *et al.* 2018).

Other areas of potential SNP typing for external visible characteristics include the identification of genetic markers that infer height, hair shaft shape (Medland *et al.* 2009; Adhikari *et al.* 2016), facial features (Liu *et al.* 2012; Claes *et al.* 2014), and male pattern baldness (Marcinska *et al.* 2015; Liu *et al.* 2016). While not specifically SNP typing, recent research into the DNA methylation patters in the human genome (epigenetics) have shown an association with age, and techniques are currently under development with the aim of inferring human chronological age to aid in forensic investigation (Zbieć-Piekarska *et al.* 2015; Park *et al.* 2016; Zubakov *et al.* 2016; Parson 2018).

1.3.2.2.2 Ancestry-informative SNPs

While autosomal STRs are the markers of choice for direct individual identification, their high mutation rates mean they are inappropriate markers for ancestry prediction. Y-STRs can also be applied for inferring the paternal biogeographic ancestry of unknown donors or missing persons. The Y-chromosome Haplotype Reference Database (YHRD) database (www.yhrd.com) includes haplotype data from 1221 populations across 135 countries, and, along with other published population studies, has shown a geographical discrimination of Y-STR haplotypes (Kayser *et al.* 1997; Kayser *et al.* 2006; Tofanelli *et al.* 2009; Purps *et al.* 2014). Y-STR profiles can be compared with those stored in the YHRD or other published databases in order to infer most likely geographical origin of paternal DNA (Kayser 2017). While they can be useful for inferring paternal ancestry, successfully amplifying and genotyping Y-STR amplicons can be difficult when working with challenging and degraded samples, and are not able to detect events of ancestry admixture (Phillips *et al.* 2009).

Ancestry-informative SNPs exhibit low mutation rates and remain stable over generations, making them important genetic descriptors of ancestry and population history (Frudakis *et al.* 2003; Butler 2007; Haasl & Payseur 2011). Ancestry SNPs must show low heterozygosity and high population heterogeneity, meaning their alleles should occur in contrasting frequencies across different populations in order to allow differentiation. The sequence information from individuals across major continental regions (Africa, Europe, the Middle East, Central and South Asia, East Asia, the Americas and Oceania) has been recorded and stored in online sources such as 1000 Genomes (The 1000 Genomes Project Consortium 2015) and Human Genome Diversity CEPH (HGDP-CEPH) (Cann *et al.* 2002) databases. These catalogues of genetic variation have allowed for the identification of SNPs that exist in high frequencies in one population versus all others. It is these loci that are targeted in SNP panels for differentiating between ancestral groups for population studies and forensic investigations.

Autosomal SNPs are inherited bi-parentally and so represent genetic input from both ancestral lineages. A panel of autosomal ancestry SNPs can generate a profile that indicates ancestry from a single ancestral gene pool, or can suggest admixture from different ancestral populations (Fondevila *et al.* 2013). Typing autosomal ancestry SNPs for forensic use has been demonstrated in the Madrid bombing attack in 2004 (Phillips et al. 2009). Specialist SNP typing using the SNPforID 34-plex ancestry SNP panel was conducted in order to determine the biogeographic ancestry of the donors of seven evidential samples where STR profiles were

unmatched to any suspects or databases. The ancestry assignment indicated a donor of North African descent, which eventually assisted in the arrest of an Algerian perpetrator (Phillips *et al.* 2009).

MtDNA SNPs are also helpful for the inference of maternal ancestry. The mtDNA coding region has observed mutation rates lower than that of the hypervariable control regions (Horai & Hayasaka 1990; Coble *et al.* 2004; Jobling *et al.* 2004). This relatively higher genetic stability of the coding region has thus been targeted for not only lineage testing but inferring the maternal ancestry of an individual. MtDNA haplogroup (a group of related haplotypes) distributions have been found to be geographically restricted historically, and can therefore help to trace a lineage back to a geographical location (Coble *et al.* 2004; Behar *et al.* 2007; Zietkiewicz *et al.* 2012). The same is true for Y-SNPs in the non-recombining Y-chromosome (NRY). The Y Chromosome Consortium (YCC) has published a database and phylogenetic tree linking 311 haplogroups in 20 major lineages (Y Chromosome Consortium 2002; Karafet *et al.* 2008) and is continually updated in accordance with emerging population studies. These Y-chromosome (Y-chr) macrohaplogroups show continental affiliations as well as more specific geographical distributions (sub-haplogroups), reflecting recent human migration (Jobling 2001; van Oven *et al.* 2011a).

Both mtDNA and Y-chr will only test for one ancestral lineage each owing to their uniparental inheritance. While often overlooked for inferring ancestry, X-chromosome SNPs also display patterns of population divergence and can provide additional information in males where the ancestry of the maternal lineage can also be deduced alongside mtDNA (Phillips 2015a; Santos *et al.* 2016). In contrast, autosomal ancestry SNPs are inherited with recombination from both the mother and father, which is especially useful for identifying individuals with admixture from different ancestral populations. Several panels exist for ancestry testing using autosomal SNPs (Fondevila *et al.* 2013; Nievergelt *et al.* 2013; Kidd *et al.* 2014; de la Puente *et al.* 2016), mtDNA coding region SNPs (Haak *et al.* 2010; van Oven *et al.* 2011b) and Y-SNPs (Haak *et al.* 2010; van Oven *et al.* 2011a; Valverde *et al.* 2013). These panels in singular, allow an individual to be assigned to one ancestral group, or in the case of autosomal SNPs that detect ancestry admixture, originating from multiple geographic locations. However, it is especially important to emphasise that the accuracy of biogeographic ancestry predictions is strongly dependant on the choice of training set reference populations, the informative value of the markers (or choice of panel) (Rosenberg *et al.* 2003), and the computational approach to assigning ancestry classifications to a sample (Cheung *et al.* 2017).

The use of SNPs that exist on the mtDNA and Y-chr to infer ancestry can also provide supplementary lineage information to resolve ambiguous or inconclusive STR relationship results, or further resolve Y-chr and mtDNA haplotypes (Kohnemann *et al.* 2008; Vallone 2012; van Oven *et al.* 2013). Traditional forensic techniques using mtDNA and the Y chromosome for identification have relied on control region sequencing and Y-STR testing. However, sometimes these techniques can fail to discriminate between distinct maternal and paternal lineages (Just *et al.* 2011). The addition of lineage SNP testing (either mtDNA coding region, whole mtDNA genome sequencing or Y-SNPs) can help to distinguish between these lineages by adding an extra element of genetic information for increased resolution power (Chaitanya *et al.* 2015; Morales-Arce *et al.* 2017).

1.4 Limitations of mtDNA and SNPs for Forensic Investigation

The principal advantage of SNP typing is the relative ease of designing short amplicons for the retrieval of genetic information from degraded DNA. However, several limitations exist with its use in forensic identification. Some issues include limited multiplexing capacity and assay design, tolerance to damaged DNA, and a lack of representative databases for some populations.

1.4.1 Multiplex Assay Design

Perhaps the most substantial drawback to SNP typing is that the predominantly bi-allelic nature of SNPs makes them relatively weakly informative per single locus. High-resolution ancestry inference requires many SNPs to reach reasonable population differentiation power with high likelihoods (Phillips 2015b). Designing large PCR multiplexes for SNP typing is a complex and difficult task that requires a delicate balance between reaching adequate discrimination and resolution power, whilst typing the minimum number of markers possible for design simplicity. Large multiplexes with a high number of PCR primers can lead to poor amplification efficiency of some loci (Apaga *et al.* 2017; de la Puente *et al.* 2017). The design of many PCR primers which do not interact with each other, and at the same time maintain short amplicons suitable for degraded DNA can lead to SNP-typing imbalance. A high number of PCR primers can also generate non-specific amplification making optimisation of PCR multiplexes a demanding task. The most commonly used SNP typing technology utilised in forensic laboratories (SNaPshot^{TM}) is based on single base extension (SBE) of PCR amplicons. SBE products are

separated based on fragment sizes within four fluorescent dye channels via capillary electrophoresis (CE).

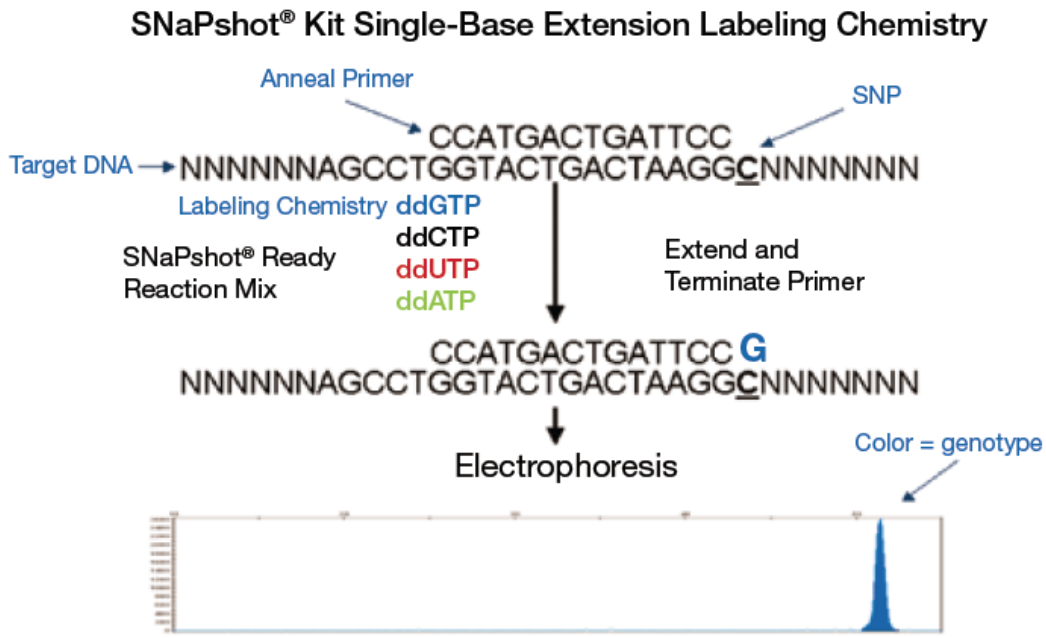


Figure 1. The SNaPshot chemistry is based on the single-base extension of unlabelled oligonucleotide or primers. Each primer binds to a complementary template in the presence of fluorescently labelled dideoxynucleotides (ddNTPs) and DNA polymerase. The polymerase extends the primer by one nucleotide, adding a single ddNTP to its 3′ end. A single peak represents a particular SNP with a known SBE fragment size, and the colour of the peak indicates which nucleotide resides at the SNP position. Taken from Applied Biosystems, 2010.

SNaPshot typing is a quick, simple and inexpensive method of interrogating small numbers of SNPs, however is limited in the number of loci that can be analysed in a single multiplex due to primer size variations and the detection chemistry of CE. Fragments with the same detected alleles can only be distinguished if they are separated by a sufficient number of nucleotides for profile readability. This proves difficult for designing large multiplexes as SNP amplicons, which can in theory be as little as 45 bp, are made much larger to accommodate the complexity with multiplex design and the fragment separation technology of SNaPshot. This of course, poses another limitation for degraded DNA, since PCR amplicons are made longer than ideal. As some SNP amplicons can exceed the upper limit of degraded DNA fragments (<200bp), primer binding sites may not be available leading to locus dropout (Zar *et al.* 2018). This may lead to poor SNP profiles or no profile at all in extremely degraded samples.

Furthermore, in order to gain a full biological profile (i.e. mtDNA SNPs, Y chromosome SNPs, autosomal SNPs for ancestry and phenotype) from degraded remains, multiple independent multiplexes need to be performed. Most forensic laboratories do not have the capacity to deliver the full range of analysis and require outsourcing, which can consume valuable DNA

extract and place a burden on time and cost resources. Y-SNPs and mtDNA SNPs are single markers that can misrepresent the overall ancestry of an individual (Phillips *et al.* 2007; Lao *et al.* 2010; Santos *et al.* 2016). The analysis of admixed ancestry (where more than one population contributes genetically to an individual) is being increasingly discussed as a concern for ancestry testing (Phillips 2015b; Cheung *et al.* 2018). Ancestry classification tests currently have high prediction accuracy for unadmixed African, European, East Asian and Oceanian populations, but have difficulty for populations with shared demographic history as a result of cross-border migration, colonisation or trade (Galanter *et al.* 2012; Phillips *et al.* 2014; de la Puente *et al.* 2017; Jin *et al.* 2018). Populations from the Americas for example, have been previously characterised as admixed due to ancient migration from East Asia, colonisation from Europe and the slave trade from Africa, and this can complicate ancestry predictions from people who originate from such a population (Galanter *et al.* 2012; Homburger *et al.* 2015). To improve accurate predictions of biogeographic ancestry, the careful selection of population-specific markers, the use of representative reference populations, appropriate classifier analysis tools and using a combination of multiple marker types has been suggested for a more comprehensive and accurate survey of the ancestry components within an individual (Cheung *et al.* 2018). The probability of misinterpreting an individual's overall ancestry using a single type of marker set is higher in places where migration from distant populations has occurred (Phillips, 2015). For these reasons, it is important to consider autosomal ancestry predictions in context of an individual's mtDNA and Y-chr (for males) haplogroup collectively in order to obtain the most accurate ancestry prediction. However, this requires a much larger set of genetic markers and multiple tests with standard SNP typing strategies using CE.

1.4.2 Lack of Representative Population Reference Databases

The strength of forensic DNA identification techniques relies heavily on the comparison of results to suitable reference population databases. A population database is a collection of allele or profile frequencies of specific genetic markers from groups of representative samples (for example ethnic groups such as African, Caucasian etc). Contrary to DNA databases (profiles from convicted criminals, persons of interest, volunteers, personal items from missing persons, and DNA profiles from unknown deceased persons), a population database is not used for matching DNA profiles, but rather is a tool for which random match probabilities and statistical tests can be performed to assess the significance of a match or result (Butler 2011). Individual identity markers, whether it be STRs or identity informative SNPs, are assessed according to

their prevalence in the population, and require statistical inference to report the strength of a match.

Australia has a strong collection of STR databases across its states. These collectively form the National Criminal Investigation DNA Database (NCIDD), a tool to help inform investigations Australia-wide. Since its inception in 2001, it now has more than 800,000 STR profiles that can be readily used to confirm the source of STR profiles recovered from crime scene samples (Australian Criminal Intelligence Commission 2016). When an STR profile from an unknown donor is consistent with that of a known person of interest or database profile, statistical likelihood tests (i.e. the likelihood that two unrelated people share the same STR profile in a given population) are performed using population databases (Taylor *et al.* 2017) in order to evaluate the evidentiary value of the match. Apart from STR markers, Australia has a very limited resource of population databases for other genetic markers.

Biogeographic ancestry can be difficult to assign to a sample from populations which previously have not been examined for ancestry-informative SNPs, or where no suitable reference population databases exist (Cheung *et al.* 2018). Although mtDNA is commonly used for familial matching and ancestry inference, no such forensic population database describing the frequency of haplotypes in the Australian population has been established. MtDNA haplotype matches are evaluated against the EMPOP mtDNA population database, which is over-represented by European populations (Parson & Dur 2007). Additionally, while Y-chr and mtDNA markers are phylogenetically informative for ancestry (Jobling & Tyler-Smith 2003), large databases with good geographical coverage are required in order to properly estimate haplotype variability in a population and thus its evidentiary value. Markers that infer autosomal ancestry have also not been examined in an Australian population to estimate the occurrence of different genetic ancestry groups. The lack of such databases can lead to interpretation issues, limiting the level of confidence that can be placed on the ancestry results and could potentially risk false assignment to country of origin for unknown remains.

The importance of choosing appropriate reference population databases has been highlighted in a previous study and guidelines for the use of mtDNA for forensic purposes (Parson & Bandelt 2007; Salas *et al.* 2007). Evaluating frequencies of mtDNA haplotypes and haplogroups can be affected significantly by the choice of source population (Parson & Bandelt 2007), even from neighbouring populations considered as closely related to the target population as possible (Salas *et al.* 2007). This can bias the frequency estimations in the target

population, and ultimately, the conclusions that are drawn from them. Therefore, population databases representative of the target population should be collated and used for the accurate interpretation of mtDNA results to avoid unreliable frequency estimates of haplotypes and haplogroups (Salas *et al.* 2007)

As already discussed, it is not uncommon in forensic laboratories to encounter DNA samples that are degraded beyond the point of detection of traditional STR testing, and so other techniques such as ancestry inference as well as mtDNA analysis are explored. Cases involving extremely degraded human DNA can include but are not limited to cold cases, missing persons and historical mass graves or war dead. In these instances, there is a limit in the confidence of DNA results as there are not only a lack of Australian population databases for those specific markers to compare to, but the results are not examined in reference to representative population databases relevant to the time period of death. Understanding the genetic composition of an historical population will allow for improved analysis and reporting of relevant cases, thus facilitating identification efforts of historical human remains.

1.5 Emerging Technologies for Degraded DNA Analysis

The advent of high-throughput, massively parallel DNA sequencing systems, collectively called Massively Parallel Sequencing (MPS) has allowed the scientific community increased knowledge of human genomic variation by permitting the contemporaneous interrogation of hundreds to thousands of genetic markers. MPS has the capability to interrogate all forensically relevant STRs and autosomal, Y-chromosome and mtDNA SNP markers in a single run without depleting extra stores of DNA extract, and overcomes a number of limitations associated with capillary electrophoresis-based SNP analysis. Equally important, whole genomes can be typed with single base-pair resolution from a limited amount of starting genetic material and has been successful even on ancient DNA samples thousands of years old (Green *et al.* 2010; Krause *et al.* 2010; Gunnarsdottir *et al.* 2011). This demonstrates the capability of these methods to retrieve genetic information from highly degraded remains with low amounts of DNA. In principle, a single sequencing run using MPS can fully replace the many different, independent forensic tests currently used since the number of targets is no longer a limitation. Targeted MPS sequencing, where only relevant markers are selected for analysis, is most applicable to forensic identification since sequencing a whole genome is sometimes not financially practical, desirable or necessary. Furthermore, the use of molecular barcoding technologies allow independent multiplexes and samples to be pooled together into a single

analysis run, such that multiple samples can be tested for hundreds of genetic markers concurrently, utilising the high throughput capacity of MPS (Binladen *et al.* 2007; Meyer & Kircher 2010; Knapp *et al.* 2012).

1.5.1 Current Commercial MPS Technologies for Human Identification

For applications to forensic DNA analysis of heavily degraded remains, MPS sequencing has a major advantage over current CE based approaches because numerous markers can be reliably sequenced simultaneously in the absence of a size separation chemistry. Methods that simultaneously type identity, ancestry, phenotype and lineage SNPs by MPS are increasingly being explored and developed. The Thermo Fisher Scientific Precision ID Panel based on the Ion Torrent platform now includes 90 autosomal SNPs and 34 Y-SNPs for individual identification (Meiklejohn & Robertson 2017). Illumina has also released the MiSeq FGx system for use with the ForenSeq™ DNA Signature Prep Kit, an MPS platform specifically designed for forensic genomics which interrogates 59 STRs, and 172 identity, phenotypic and biogeographical ancestry SNPs (Churchill *et al.* 2016). The Qiagen 140-SNP forensic identification multiplex types 140 SNPs informative for individual identity and can be sequenced on either the Ion Torrent or MiSeq platforms (de la Puente *et al.* 2017). However, such methods are still based on an initial PCR amplification of the target loci which does not eliminate PCR biases and the difficulties in designing multiplexes for a large number of markers as already discussed. Because of this, SNPs that can in principle generate amplicons as low as 45bp are in practice much larger. In some cases this exceeds 200bp, already greater than the upper limit of the average fragment length of degraded samples, and has limited success on degraded DNA below 150bp (Knapp & Hofreiter 2010; Gettings *et al.* 2015; Bulbul & Filoglu 2018). In addition, the use of a higher number of primer pairs in a single assay can lead to PCR biases, inefficiency and underperformance of some markers (Apaga *et al.* 2017; de la Puente *et al.* 2017), which makes PCR multiplex optimisation a complicated task. This can become a barrier for some forensic laboratories that require the design of tailored panels for specific purposes.

As discussed, chemical modifications to degraded DNA (e.g. blocking lesions) can cause unsuccessful PCR amplification of target amplicons, an issue that is still faced by using PCR multiplexes in MPS approaches. PCR amplification failure may also arise due to the presence of PCR inhibitors in a sample such as heme, melanin, heparin and humic substances (Elwick *et al.* 2018). The absence of any unique barcoding system for each sample prior to PCR

amplification can also be a concern for detecting and filtering laboratory contamination. Because of these reasons, although a valuable tool, current commercial MPS approaches using PCR for forensic human identification of degraded remains have some technical drawbacks.

1.5.2 Hybridisation Enrichment as an Alternative Approach for Typing Degraded DNA

Hybridisation enrichment (or hybridisation capture) for MPS relies on the binding of biotinylated DNA or RNA probes that are complementary to target regions in a DNA sample (Mertes *et al.* 2011). This strategy can enrich for SNP loci prior to sequencing without the need for an initial PCR. Streptavidin beads magnetise to probes bound to target DNA, while unbound DNA and impurities are eliminated through a series of stringency washes. Hybridisation enrichment can eliminate some issues with PCR-based approaches, particularly for primer design, and as a result much shorter fragment lengths of DNA can be captured without the need for intact PCR primer binding sites (Schubert *et al.* 2012). There is no requirement for complex PCR primer multiplex design for large numbers of markers and thus no limit on how many loci can be examined in a single assay. Some hybridisation enrichment strategies target tens of thousands of SNPs in a single panel and have been successful on ancient DNA samples (Soubrier *et al.* 2016). Recent studies using this target enrichment strategy have demonstrated success in recovering short sequences in highly degraded DNA for ancient DNA and forensic analyses (Templeton *et al.* 2013; Soubrier *et al.* 2016). Templeton *et al.* (2013) were able to recover whole mitochondrial genomes with high resolution and sequencing coverage and depth on samples that had previously failed (or was of low-resolution) with standard forensic identity testing. Another more recent study applied a custom hybridisation enrichment panel including 307 SNPs and 36 microhaplotypes in the nuclear genome, with a focus on identity informative SNPs (Bose *et al.* 2018). Other markers in this panel include those for phenotype and ancestry, and 70 tri- and tetra-allelic SNPs for mixture resolution (Bose *et al.* 2018). Both studies have explored the application of this relatively recent hybridisation enrichment strategy for forensics purposes with good success.

The advent of MPS has undoubtedly sparked a movement towards generating massive amounts of data from biological samples. While an exciting age for DNA analysis, forensic investigations involving DNA analysis must still be focused on retrieving genetic information relevant to specific forensic questions. Current forensic panels using MPS technologies have demonstrated utility for generating data for hundreds of markers from forensic-type samples (de la Puente *et al.* 2017; Meiklejohn & Robertson 2017; Xavier & Parson 2017), yet present

an inflexible approach to SNP typing. The SNPs targeted in these particular panels are interrogated by pre-mixed primer sets or probe pools that could result in an excess of data impertinent to the questions of interest. The concerns over the generation of excess data are increasingly being explored in the forensic community after recognising the possibilities over developing MPS technologies as a potential invasion of 'genetic privacy' (Scudder *et al.* 2018b, 2018a). The generation of 'big data' also raises concerns over the extra demand for interpretation, computing power and bioinformatic expertise for such larger bodies of genetic data per sample, and how this should be managed in a forensic laboratory (Phillips 2018; Scudder *et al.* 2018b).

1.6 Overview of Thesis and Data Chapter Summaries

Collectively, this thesis aims to develop and explore alternative DNA intelligence methods for human identification of highly degraded and historical Australian remains. The following manuscripts have been compiled to explore several laboratory and analysis methods to improve the genotyping of degraded human DNA for forensic intelligence purposes. New data generated from a sample of individuals representing an Australian historical population will help inform the ancestry analysis of historical Australian remains. From the knowledge gained throughout this project, new workflows and considerations will be proposed and explored.

Chapter 2: A mini-multiplex SNaPshot assay for the triage of degraded human DNA
The first study aimed to develop a novel triaging tool based on SNP typing for the purpose of screening DNA samples for DNA quality and broad biological profile prior to deciding on which downstream laboratory process are most likely to retrieve sufficient genetic data for analysis. The newly developed panel interrogates 18 SNP and indel markers across nuclear and mtDNA targets with varying amplicon sizes to qualitatively assess DNA degradation and to triage priority of DNA samples based on inferred sex, mtDNA and Y-chromosome ancestry, and eye colour. Firstly, I apply the panel to a set of reference samples with known biogeographic ancestry, sex and eye colour to establish the accuracy and any interpretation considerations of the panel. I then demonstrate the utility of the method for SNP retrieval, and the inference of ancestry, sex and phenotype on a range of degraded human DNA extracts and discuss the value of such a tool in forensic analysis of degraded DNA.

Chapter 3: A custom hybridisation enrichment forensic intelligence panel to infer biogeographic ancestry, hair and eye colour, and Y chromosome lineage.

New hybridisation enrichment strategies for MPS analysis of degraded and trace DNA samples have increased the capacity to retrieve genetic data for forensic identification and intelligence gathering (Templeton *et al.* 2013; Bose *et al.* 2018; Shih *et al.* 2018). Chapter 3 describes the SNP selection and methods development of a novel 124-SNP panel for biogeographic ancestry, paternal lineage, and hair and eye colour inference. An evaluation of the panel is presented to explore whether the inferences made from the custom panel of SNP markers are robust and accurate for predicting a range of biogeographic ancestries, sex, phenotype and Y-chr haplogroups. This was performed on a range of samples of known sex, self-declared ancestry and hair and eye colour. The study explores the value and feasibility of the panel to aid in intelligence gathering for forensic investigation.

Chapter 4: Application of the Miniplex SNaPshot assay and the 124-SNP hybridisation enrichment assay to degraded human DNA

Chapter 4 brings together the two newly developed and applied SNP typing tools, the Miniplex from Chapter 2, and the custom enrichment panel from Chapter 3 on a range of degraded human teeth and forensic casework samples to demonstrate real-world applications of the methods. The successes and limitations of the combined analysis workflows are discussed, and further suggestions are proposed for possible avenues for improvement of the techniques in obtaining genetic data for inference of forensically relevant intelligence information.

Chapter 5: The Historical Australian DNA Database

Current ancestry testing of historical Australian remains suffers from a lack of a representative population database during the early 20th century. The biogeographic ancestry composition of the Australian population during this time is therefore largely unknown, and this may impact on the accuracy and reliability of the sorting of recovered historical remains based on genetic ancestry. The final study of this thesis aimed to collate the first Australian historical population database that will continue to be made larger with further study. DNA samples were collected from members of the public who reflect the Australian population prior to the waves of migration into Australia after 1945. Their biogeographic ancestry was determined through mtDNA analysis and by the use of an autosomal ancestry SNP panel to detect and characterise the biogeographic ancestry composition of the Australian population during this time. This approach has produced new ancestry data and generates the

foundations of the first multi-gene historical DNA database for Australia and describes its value for evaluating ancestry prediction results from recovered human remains.

References

Adhikari, K., Fontanil, T., Cal, S., Mendoza-Revilla, J., Fuentes-Guajardo, M., Chacón-Duque, J.-C., Al-Saadi, F., Johansson, J.A., Quinto-Sanchez, M., Acuña-Alonzo, V., et al. 2016. A genome-wide association scan in admixed Latin Americans identifies loci influencing facial and scalp hair features, *Nat Commun*, 7, 10815.

Akkoc, R. 2014, '2004 Boxing Day tsunami facts. ', *The Telegraph*, 19 December, 2014, <http://www.telegraph.co.uk/news/worldnews/asia/11303114/2004-Boxing-Day-tsunami-facts.html. >

Alaeddini, R., Walsh, S.J. & Abbas, A. 2010. Forensic implications of genetic analyses from degraded DNA--a review, *Forensic Sci Int Genet*, 4, 148-57.

Allard, M.W., Miller, K., Wilson, M., Monson, K. & Budowle, B. 2002. Characterization of the Caucasian haplogroups present in the SWGDAM forensic mtDNA dataset for 1771 human control region sequences. Scientific Working Group on DNA Analysis Methods, *J Forensic Sci*, 47, 1215-23.

Alonso, A., Martin, P., Albarran, C., Garcia, P., Fernandez de Simon, L., Jesus Iturralde, M., Fernandez-Rodriguez, A., Atienza, I., Capilla, J., Garcia-Hirschfeld, J., et al. 2005. Challenges of DNA profiling in mass disaster investigations, *Croat Med J*, 46, 540-8.

Amorim, A. & Pereira, L. 2005. Pros and cons in the use of SNPs in forensic kinship investigation: a comparative analysis with STRs, *Forensic Sci Int*, 150, 17-21.

Apaga, D.L.T., Dennis, S.E., Salvador, J.M., Calacal, G.C. & De Ungria, M.C.A. 2017. Comparison of Two Massively Parallel Sequencing Platforms using 83 Single Nucleotide Polymorphisms for Human Identification, *Sci Rep*, 7, 398.

Applied Biosystems 2016, 'GlobalFiler™ PCR Amplification Kit: User Guide. ', <https://www.thermofisher.com/order/catalog/product/4476135>.

Australian Criminal Intelligence Commission 2016, *National Criminal Investigation DNA Database*, viewed August 10, 2017, <Available at https://www.acic.gov.au/our-services/biometric-matching/national-criminal-investigation-dna-database >.

Barbaro, A., Cormaci, P. & Barbaro, A. 2006. X-STR typing for an identification casework, *International Congress Series*, 1288, 513-5.

Behar, D.M., Rosset, S., Blue-Smith, J., Balanovsky, O., Tzur, S., Comas, D., Mitchell, R.J., Quintana-Murci, L., Tyler-Smith, C., Wells, R.S., et al. 2007. The Genographic Project Public Participation Mitochondrial DNA Database, *PLoS Genet*, 3, e104.

Binladen, J., Gilbert, M.T.P., Bollback, J.P., Panitz, F., Bendixen, C., Nielsen, R. & Willerslev, E. 2007. The Use of Coded PCR Primers Enables High-Throughput Sequencing of Multiple Homolog Amplification Products by 454 Parallel Sequencing, *PLoS One*, 2, e197.

Bogas, V., Carvalho, M., Anjos, M.J. & Corte-Real, F. 2016. Genetic identification of degraded and/or inhibited DNA samples, *Aust J Forensic Sci*, 48, 381-406.

Bogas, V., Carvalho, M., Corte-Real, F. & Porto, M.J. 2015. Testing the behavior of GlobalFiler® PCR amplification kit with degraded and/or inhibited biological samples, *Forensic Sci Int Genet Supp Series*, 5, e21-e3.

Bose, N., Carlberg, K., Sensabaugh, G., Erlich, H. & Calloway, C. 2018. Target capture enrichment of nuclear SNP markers for massively parallel sequencing of degraded and mixed samples, *Forensic Sci Int Genet*, 34, 186-96.

Bricknell, S. 2017, *Missing persons: Who is at risk?*, Australian Institute of Criminology, Canberra, Australia.

Brinkmann, B., Klintschar, M., Neuhuber, F., Hühne, J. & Rolf, B. 1998. Mutation Rate in Human Microsatellites: Influence of the Structure and Length of the Tandem Repeat, *The American Journal of Human Genetics*, 62, 1408-15.

Budowle, B. 2004. SNP typing strategies, *Forensic Sci Int*, 146 Suppl, S139-42.

Budowle, B., Allard, M.W., Wilson, M.R. & Chakraborty, R. 2003. Forensics and mitochondrial DNA: applications, debates, and foundations, *Annu Rev Genomics Hum Genet*, 4, 119-41.

Budowle, B. & van Daal, A. 2008. Forensically relevant SNP classes, *Biotechniques*, 44, 603-8, 10.

Budowle, B., Wilson, M.R., DiZinno, J.A., Stauffer, C., Fasano, M.A., Holland, M.M. & Monson, K.L. 1999. Mitochondrial DNA regions HVI and HVII population data, *Forensic Sci Int*, 103, 23-35.

Bulbul, O. & Filoglu, G. 2018. Development of a SNP panel for predicting biogeographical ancestry and phenotype using massively parallel sequencing, *Electrophoresis*, 39, 2743-51.

Burger, J., Hummel, S., Hermann, B. & Henke, W. 1999. DNA preservation: a microsatellite-DNA study on ancient skeletal remains, *Electrophoresis*, 20, 1722-8.

Butler, J.M. 2005. *Forensic DNA Typing: Biology, Technology, and Genetics of STR Markers.*, 2nd edn, Elsevier Academic Press.

Butler, J.M. 2007. Short tandem repeat typing technologies used in human identity testing, *Biotechniques*, 43, ii-v.

Butler, J.M. 2009. *Fundamentals of Forensic DNA Typing*, Academic Press, London.

Butler, J.M. 2011. *Advanced Topics in Forensic DNA Typing: Methodology*, Academic, London.

Cann, H.M., de Toma, C., Cazes, L., Legrand, M.F., Morel, V., Piouffre, L., Bodmer, J., Bodmer, W.F., Bonne-Tamir, B., Cambon-Thomsen, A., et al. 2002. A human genome diversity cell line panel, *Science*, 296, 261-2.

Chaitanya, L., Breslin, K., Zuñiga, S., Wirken, L., Pośpiech, E., Kukla-Bartoszek, M., Sijen, T., Knijff, P.d., Liu, F., Branicki, W., et al. 2018. The HIrisPlex-S system for eye, hair and skin colour prediction from DNA: Introduction and forensic developmental validation, *Forensic Sci Int Genet*, 35, 123-35.

Chaitanya, L., Pajnič, I.Z., Walsh, S., Balažic, J., Zupanc, T. & Kayser, M. 2017. Bringing colour back after 70 years: Predicting eye and hair colour from skeletal remains of World War II victims using the HIrisPlex system, *Forensic Sci Int Genet*, 26, 48-57.

Chaitanya, L., Ralf, A., van Oven, M., Kupiec, T., Chang, J., Lagacé, R. & Kayser, M. 2015. Simultaneous Whole Mitochondrial Genome Sequencing with Short Overlapping Amplicons Suitable for Degraded DNA Using the Ion Torrent Personal Genome Machine, *Human Mutation*, 36, 1236-47.

Cheung, E.Y.Y., Gahan, M.E. & McNevin, D. 2017. Prediction of biogeographical ancestry from genotype: a comparison of classifiers, *Int J Legal Med*, 131, 901-12.

Cheung, E.Y.Y., Gahan, M.E. & McNevin, D. 2018. Prediction of biogeographical ancestry in admixed individuals, *Forensic Sci Int Genet*, 36, 104-11.

Churchill, J.D., Schmedes, S.E., King, J.L. & Budowle, B. 2016. Evaluation of the Illumina® Beta Version ForenSeq™ DNA Signature Prep Kit for use in genetic profiling, *Forensic Sci Int Genet*, 20, 20-9.

Claes, P., Liberton, D.K., Daniels, K., Rosana, K.M., Quillen, E.E., Pearson, L.N., McEvoy, B., Bauchet, M., Zaidi, A.A., Yao, W., et al. 2014. Modeling 3D Facial Shape from DNA, *PLoS Genet*, 10, e1004224.

Coble, M.D., Just, R.S., O'Callaghan, J.E., Letmanyi, I.H., Peterson, C.T., Irwin, J.A. & Parsons, T.J. 2004. Single nucleotide polymorphisms over the entire mtDNA genome that increase the power of forensic testing in Caucasians, *Int J Legal Med*, 118.

Coble, M.D., Loreille, O.M., Wadhams, M.J., Edson, S.M., Maynard, K., Meyer, C.E., Niederstätter, H., Berger, C., Berger, B., Falsetti, A.B., et al. 2009. Mystery Solved: The Identification of the Two Missing Romanov Children Using DNA Analysis, *PLoS One*, 4, e4838.

Collins, F.S., Brooks, L.D. & Chakravarti, A. 1998. A DNA polymorphism discovery resource for research on human genetic variation, *Genome Res*, 8, 1229-31.

Crainic, K., Paraire, F., Leterreux, M., Durigon, M. & de Mazancourt, P. 2002. Skeletal remains presumed submerged in water for three years identified using PCR-STR analysis, *J Forensic Sci*, 47, 1025-7.

Dabney, J., Meyer, M. & Paabo, S. 2013. Ancient DNA damage, *Cold Spring Harb Perspect Biol*, 5.

de la Puente, M., Phillips, C., Santos, C., Fondevila, M., Carracedo, Á. & Lareu, M.V. 2017. Evaluation of the Qiagen 140-SNP forensic identification multiplex for massively parallel sequencing, *Forensic Sci Int Genet*, 28, 35-43.

de la Puente, M., Santos, C., Fondevila, M., Manzo, L., Carracedo, A., Lareu, M.V. & Phillips, C. 2016. The Global AIMs Nano set: A 31-plex SNaPshot assay of ancestry-informative SNPs, *Forensic Sci Int Genet*, 22, 81-8.

Dixon, L.A., Murray, C.M., Archer, E.J., Dobbins, A.E., Koumi, P. & Gill, P. 2005. Validation of a 21-locus autosomal SNP multiplex for forensic identification purposes, *Forensic Sci Int*, 154, 62-77.

Edson, S.M., Ross, J.P., Coble, M.D., Parsons, T.J. & Barritt, S.M. 2004. Naming the Dead — Confronting the Realities of Rapid Identification of Degraded Skeletal Remains, *Forensic Sci Rev*, 16, 63.

Elwick, K., Zeng, X., King, J., Budowle, B. & Hughes-Stamm, S. 2018. Comparative tolerance of two massively parallel sequencing systems to common PCR inhibitors, *Int J Legal Med*, 132, 983-95.

Fondevila, M., Phillips, C., Naverán, N., Cerezo, M., Rodríguez, A., Calvo, R., Fernández, L.M., Carracedo, Á. & Lareu, M.V. 2008a. Challenging DNA: Assessment of a range of genotyping approaches for highly degraded forensic samples, *Forensic Sci Int Genet Supp Series*, 1, 26-8.

Fondevila, M., Phillips, C., Naveran, N., Fernandez, L., Cerezo, M., Salas, A., Carracedo, A. & Lareu, M.V. 2008b. Case report: identification of skeletal remains using short-amplicon marker analysis of severely degraded DNA extracted from a decomposed and charred femur, *Forensic Sci Int Genet*, 2, 212-8.

Fondevila, M., Phillips, C., Santos, C., Freire Aradas, A., Vallone, P.M., Butler, J.M., Lareu, M.V. & Carracedo, Á. 2013. Revision of the SNPforID 34-plex forensic ancestry test: Assay enhancements, standard reference sample genotypes and extended population studies, *Forensic Sci Int Genet*, 7, 63-74.

Foran, D.R. 2006. Relative Degradation of Nuclear and Mitochondrial DNA: An Experimental Approach*, *J Forensic Sci*, 51, 766-70.

Freire-Aradas, A., Fondevila, M., Kriegel, A.K., Phillips, C., Gill, P., Prieto, L., Schneider, P.M., Carracedo, Á. & Lareu, M.V. 2012. A new SNP assay for identification of highly degraded human DNA, *Forensic Sci Int Genet*, 6, 341-9.

Fridman, C., Cardena, M.M.S.G., Kanto, E.A., Godinho, M.B.C. & Gonçalves, F.T. 2011. SNPs in mitochondrial DNA coding region used to discriminate common sequences in HV1–HV2–HV3 region, *Forensic Sci Int Genet Supp Series*, 3, e75-e6.

Frudakis, T., Venkateswarlu, K., Thomas, M.J., Gaskin, Z., Ginjupalli, S., Gunturi, S., Ponnuswamy, V., Natarajan, S. & Nachimuthu, P.K. 2003. A classifier for the SNP-based inference of ancestry, *J Forensic Sci*, 48, 771-82.

Galanter, J.M., Fernandez-Lopez, J.C., Gignoux, C.R., Barnholtz-Sloan, J., Fernandez-Rozadilla, C., Via, M., Hidalgo-Miranda, A., Contreras, A.V., Figueroa, L.U., Raska, P., et al. 2012. Development of a Panel of Genome-Wide Ancestry Informative Markers to Study Admixture Throughout the Americas, *PLoS Genet*, 8, e1002554.

Gettings, K.B., Kiesler, K.M. & Vallone, P.M. 2015. Performance of a next generation sequencing SNP assay on degraded DNA, *Forensic Sci Int Genet*, 19, 1-9.

Gill, P. 2001. An assessment of the utility of single nucleotide polymorphisms (SNPs) for forensic purposes, *Int J Legal Med*, 114, 204-10.

Goodsell, D.S. 2001. The molecular perspective: ultraviolet light and pyrimidine dimers, *Oncologist*, 6, 298-9.

Gopinath, S., Zhong, C., Nguyen, V., Ge, J., Lagace, R.E., Short, M.L. & Mulero, J.J. 2016. Developmental validation of the Yfiler® Plus PCR Amplification Kit: An enhanced Y-STR multiplex for casework and database applications, *Forensic Sci Int Genet*, 24, 164-75.

Green, R.E., Krause, J., Briggs, A.W., Maricic, T., Stenzel, U., Kircher, M., Patterson, N., Li, H., Zhai, W., Fritz, M.H.-Y., et al. 2010. A Draft Sequence of the Neandertal Genome, *Science*, 328, 710-22.

Gunnarsdottir, E.D., Li, M., Bauchet, M., Finstermeier, K. & Stoneking, M. 2011. High-throughput sequencing of complete human mtDNA genomes from the Philippines, *Genome Res*, 21, 1-11.

Haak, W., Balanovsky, O., Sanchez, J.J., Koshel, S., Zaporozhchenko, V., Adler, C.J., der Sarkissian, C., Brandt, G., Schwarz, C., Nicklisch, N., et al. 2010. Ancient DNA from European early Neolithic farmers reveals their near eastern affinities, *PLoS Biol*, 8.

Haasl, R.J. & Payseur, B.A. 2011. Multi-locus inference of population structure: a comparison between single nucleotide polymorphisms and microsatellites, *Heredity*, 106, 158-71.

Han, J., Kraft, P., Nan, H., Guo, Q., Chen, C., Qureshi, A., Hankinson, S.E., Hu, F.B., Duffy, D.L., Zhao, Z.Z., et al. 2008. A genome-wide association study identifies novel alleles associated with hair color and skin pigmentation, *PLoS Genet*, 4, e1000074.

Hartman, D., Benton, L., Spiden, M. & Stock, A. 2015. The Victorian missing persons DNA database – two interesting case studies, *Aust J Forensic Sci*, 47, 161-72.

Higgins, D., Rohrlach, A.B., Kaidonis, J., Townsend, G. & Austin, J.J. 2015. Differential Nuclear and Mitochondrial DNA Preservation in Post-Mortem Teeth with Implications for Forensic and Ancient DNA Studies, *PLoS One*, 10, e0126935.

Holland, M.M., Fisher, D.L., Mitchell, L.G., Rodriquez, W.C., Canik, J.J., Merril, C.R. & Weedn, V.W. 1993. Mitochondrial DNA sequence analysis of human skeletal remains: identification of remains from the Vietnam War, *J Forensic Sci*, 38, 542-53.

Homburger, J.R., Moreno-Estrada, A., Gignoux, C.R., Nelson, D., Sanchez, E., Ortiz-Tello, P., Pons-Estel, B.A., Acevedo-Vasquez, E., Miranda, P., Langefeld, C.D., et al. 2015. Genomic Insights into the Ancestry and Demographic History of South America, *PLoS Genet*, 11, e1005602.

Horai, S. & Hayasaka, K. 1990. Intraspecific nucleotide sequence differences in the major noncoding region of human mitochondrial DNA, *Am J Hum Genet*, 46, 828-42.

Israr, M., Rafiq, S., Shahid, A.A., Hussain, H. & Rakha, A. 2014. Scope of X-Chromosomal MiniSTRs: Current Developments, *J Forensic Res*, 5.

Jin, S., Chase, M., Henry, M., Alderson, G., Morrow, J.M., Malik, S., Ballard, D., McGrory, J., Fernandopulle, N., Millman, J., et al. 2018. Implementing a biogeographic ancestry inference service for forensic casework, *Electrophoresis*.

Jobling, M. 2001. Y-chromosomal SNP haplotype diversity in forensic analysis, *Forensic Sci Int*, 118, 158-62.

Jobling, M., Hurles, M. & Tyler-Smith, C. 2004. *Human Evolutionary Genetics: Origins, Peoples & Disease*, Garland Science, New York.

Jobling, M. & Tyler-Smith, C. 2003. The human Y chromosome: an evolutionary marker comes of age, *Nat Rev Genet*, 4, 598-612.

Just, R.S., Loreille, O.M., Molto, J.E., Merriwether, D.A., Woodward, S.R., Matheson, C., Creed, J., McGrath, S.E., Sturk-Andreaggi, K., Coble, M.D., et al. 2011. Titanic's unknown child: The critical role of the mitochondrial DNA coding region in a re-identification effort, *Forensic Sci Int Genet*, 5, 231-5.

Karafet, T.M., Mendez, F.L., Meilerman, M.B., Underhill, P.A., Zegura, S.L. & Hammer, M.F. 2008. New binary polymorphisms reshape and increase resolution of the human Y chromosomal haplogroup tree, *Genome Res*, 18, 830-8.

Kayser, M. 2017. Forensic use of Y-chromosome DNA: a general overview, *Human Genetics*, 136, 621-35.

Kayser, M., Brauer, S., Cordaux, R., Casto, A., Lao, O., Zhivotovsky, L.A., Moyse-Faurie, C., Rutledge, R.B., Schiefenhoevel, W., Gil, D., et al. 2006. Melanesian and Asian Origins of Polynesians: mtDNA and Y Chromosome Gradients Across the Pacific, *Mol Biol Evol*, 23, 2234-44.

Kayser, M., Caglià, A., Corach, D., Fretwell, N., Gehrig, C., Graziosi, G., Heidorn, F., Herrmann, S., Herzog, B., Hidding, M., et al. 1997. Evaluation of Y-chromosomal STRs: a multicenter study, *Int J Legal Med*, 110, 125-33.

Kayser, M. & de Knijff, P. 2011. Improving human forensics through advances in genetics, genomics and molecular biology, *Nat Rev Genet*, 12, 179-92.

Kidd, K.K., Speed, W.C., Pakstis, A.J., Furtado, M.R., Fang, R., Madbouly, A., Maiers, M., Middha, M., Friedlaender, F.R. & Kidd, J.R. 2014. Progress toward an efficient panel of SNPs for ancestry inference, *Forensic Sci Int Genet*, 10, 23-32.

Kline, M.C., Duewer, D.L., Redman, J.W. & Butler, J.M. 2005. Results from the NIST 2004 DNA quantitation study, *J Forensic Sci*, 50, 571-8.

Knapp, M. & Hofreiter, M. 2010. Next Generation Sequencing of Ancient DNA: Requirements, Strategies and Perspectives, *Genes*, 1, 227-43.

Knapp, M., Stiller, M. & Meyer, M. 2012, 'Generating barcoded libraries for multiplex high-throughput sequencing', in DNA Ancient (ed.), *Edited by Shapiro B, Hofreiter M*, New York, Humana Press.

Kohnemann, S., Sibbing, U., Pfeiffer, H. & Hohoff, C. 2008. A rapid mtDNA assay of 22 SNPs in one multiplex reaction increases the power of forensic testing in European Caucasians, *Int J Legal Med*, 122, 517-23.

Krause, J., Fu, Q., Good, J.M., Viola, B., Shunkov, M.V., Derevianko, A.P. & Pääbo, S. 2010. The complete mitochondrial DNA genome of an unknown hominin from southern Siberia, *Nature*, 464, 894.

Lao, O., Vallone, P.M., Coble, M.D., Diegoli, T.M., van Oven, M., van der Gaag, K.J., Pijpe, J., de Knijff, P. & Kayser, M. 2010. Evaluating Self-declared Ancestry of U.S. Americans with Autosomal, Y-chromosomal and Mitochondrial DNA, *Human Mutat*, 31, e1875-e93.

Legros, F., Malka, F., Frachon, P., Lombès, A. & Rojo, M. 2004. Organization and dynamics of human mitochondrial DNA, *J Cell Sci*, 117, 2653.

Levy-Booth, D.J., Campbell, R.G., Gulden, R.H., Hart, M.M., Powell, J.R., Klironomos, J.N., Pauls, K.P., Swanton, C.J., Trevors, J.T. & Dunfield, K.E. 2007. Cycling of extracellular DNA in the soil environment, *Soil Biol and Biochem*, 39, 2977-91.

Lindahl, T. 1993. Instability and decay of the primary structure of DNA, *Nature*, 362, 709.

Liu, F., Hamer, M.A., Heilmann, S., Herold, C., Moebus, S., Hofman, A., Uitterlinden, A.G., Nöthen, M.M., van Duijn, C.M., Nijsten, T.E., et al. 2016. Prediction of male-pattern baldness from genotypes, *Eur J Hum Genet*, 24, 895-902.

Liu, F., van der Lijn, F., Schurmann, C., Zhu, G., Chakravarty, M.M., Hysi, P.G., Wollstein, A., Lao, O., de Bruijne, M., Ikram, M.A., et al. 2012. A Genome-Wide Association Study Identifies Five Loci Influencing Facial Morphology in Europeans, *PLoS Genet*, 8, e1002932.

Marcinska, M., Pospiech, E., Abidi, S., Andersen, J.D., van den Berge, M., Carracedo, A., Eduardoff, M., Marczakiewicz-Lustig, A., Morling, N., Sijen, T., et al. 2015. Evaluation of DNA variants associated with androgenetic alopecia and their potential to predict male pattern baldness, *PLoS One*, 10, e0127852.

Matheson, C., Gurney, C., Esau, N. & Lehto, R. 2010. Assessing PCR Inhibition from Humic Substances, *Open Enzym Inhib J*, 3, 38-45.

Mattayat, D., Kitpipit, T., Phetpeng, S., Asawutmangkul, W. & Thanakiatkrai, P. 2016. Comparative performance of AmpFLSTR® Identifiler® Plus PCR amplification kit and QIAGEN® Investigator® IDplex Plus kit, *Sci Justice*, 56, 468-74.

Medland, S.E., Zhu, G. & Martin, N.G. 2009. Estimating the heritability of hair curliness in twins of European ancestry, *Twin Res Hum Genet*, 12, 514-8.

Meiklejohn, K.A. & Robertson, J.M. 2017. Evaluation of the Precision ID Identity Panel for the Ion Torrent™ PGM™ sequencer, *Forensic Sci Int Genet*, 31, 48-56.

Mertes, F., Elsharawy, A., Sauer, S., van Helvoort, J.M., van der Zaag, P.J., Franke, A., Nilsson, M., Lehrach, H. & Brookes, A.J. 2011. Targeted enrichment of genomic DNA regions for next-generation sequencing, *Brief Funct Genomics*, 10, 374-86.

Meyer, M. & Kircher, M. 2010. Illumina sequencing library preparation for highly multiplexed target capture and sequencing, *Cold Spring Harb Protoc*, 2010, pdb.prot5448.

Morales-Arce, A.Y., Hofman, C.A., Duggan, A.T., Benfer, A.K., Katzenberg, M.A., McCafferty, G. & Warinner, C. 2017. Successful reconstruction of whole mitochondrial genomes from ancient Central America and Mexico, *Sci Rep*, 7, 18100.

Mulero, J.J., Chang, C.W., Lagace, R.E., Wang, D.Y., Bas, J.L., McMahon, T.P. & Hennessy, L.K. 2008. Development and validation of the AmpFlSTR MiniFiler PCR Amplification Kit: a MiniSTR multiplex for the analysis of degraded and/or PCR inhibited DNA, *J Forensic Sci*, 53, 838-52.

Musgrave-Brown, E., Ballard, D., Balogh, K., Bender, K., Berger, B., Bogus, M., Børsting, C., Brion, M., Fondevila, M., Harrison, C., et al. 2007. Forensic validation of the SNPforID 52-plex assay, *Forensic Sci Int Genet*, 1, 186-90.

Nelson, J. 2009, *Repair of damaged DNA for forensic analysis*, GE Global Research Centre, Niskayuna, New York.

Nievergelt, C.M., Maihofer, A.X., Shekhtman, T., Libiger, O., Wang, X., Kidd, K.K. & Kidd, J.R. 2013. Inference of human continental origin and admixture proportions using a highly discriminative ancestry informative 41-SNP panel, *Investig Genet*, 4, 13.

Opel, K.L., Chung, D.T., Drabek, J., Tatarek, N.E., Jantz, L.M. & McCord, B.R. 2006. The application of miniplex primer sets in the analysis of degraded DNA from human skeletal remains, *J Forensic Sci*, 51, 351-6.

Pakstis, A.J., Speed, W.C., Fang, R., Hyland, F.C., Furtado, M.R., Kidd, J.R. & Kidd, K.K. 2010. SNPs for a universal individual identification panel, *Hum Genet*, 127, 315-24.

Park, J.-L., Kim, J.H., Seo, E., Bae, D.H., Kim, S.-Y., Lee, H.-C., Woo, K.-M. & Kim, Y.S. 2016. Identification and evaluation of age-correlated DNA methylation markers for forensic use, *Forensic Sci Int Genet*, 23, 64-70.

Parson, W. 2018. Age Estimation with DNA: From Forensic DNA Fingerprinting to Forensic (Epi)Genomics: A Mini-Review, *Gerontology*, 64, 326-32.

Parson, W. & Bandelt, H.J. 2007. Extended guidelines for mtDNA typing of population data in forensic science, *Forensic Sci Int Genet*, 1, 13-9.

Parson, W. & Dur, A. 2007. EMPOP--a forensic mtDNA database, *Forensic Sci Int Genet*, 1, 88-92.

Parsons, T.J. & Coble, M.D. 2001. Increasing the forensic discrimination of mitochondrial DNA testing through analysis of the entire mitochondrial DNA genome, *Croat Med J*, 42, 304-9.

Phillips, C. 2015a, 'Ancestry Informative Markers', in MM Houck (ed.), *Forensic Biology*, 1st edn, Academic Press, Waltham, pp. 125-34.

Phillips, C. 2015b. Forensic genetic analysis of bio-geographical ancestry, *Forensic Sci Int Genet*, 18, 49-65.

Phillips, C. 2018. The Golden State Killer investigation and the nascent field of forensic genealogy, *Forensic Sci Int Genet*, 36, 186-8.

Phillips, C., Lareu, M., Sanchez, J., Brion, M., Sobrino, B., Morling, N., Schneider, P., Syndercombe Court, D. & Carracedo, A. 2004. Selecting single nucleotide polymorphisms for forensic applications, *International Congress Series*, 1261, 18-20.

Phillips, C., Parson, W., Lundsberg, B., Santos, C., Freire-Aradas, A., Torres, M., Eduardoff, M., Børsting, C., Johansen, P., Fondevila, M., et al. 2014. Building a forensic ancestry panel from the ground up: The EUROFORGEN Global AIM-SNP set, *Forensic Sci Int Genet*, 11, 13-25.

Phillips, C., Prieto, L., Fondevila, M., Salas, A., Gómez-Tato, A., Álvarez-Dios, J., Alonso, A., Blanco-Verea, A., Brión, M., Montesino, M., et al. 2009. Ancestry Analysis in the 11-M Madrid Bomb Attack Investigation, *PLoS One*, 4, e6583.

Phillips, C., Salas, A., Sanchez, J.J., Fondevila, M., Gomez-Tato, A., Alvarez-Dios, J., Calaza, M., de Cal, M.C., Ballard, D., Lareu, M.V., et al. 2007. Inferring ancestral origin using a single multiplex assay of ancestry-informative marker SNPs, *Forensic Sci Int Genet*, 1, 273-80.

Purps, J., Siegert, S., Willuweit, S., Nagy, M., Alves, C., Salazar, R., Angustia, S.M.T., Santos, L.H., Anslinger, K., Bayer, B., et al. 2014. A global analysis of Y-chromosomal haplotype diversity for 23 STR loci, *Forensic Sci Int Genet*, 12, 12-23.

Quintans, B., Alvarez-Iglesias, V., Salas, A., Phillips, C., Lareu, M.V. & Carracedo, A. 2004. Typing of mitochondrial DNA coding region SNPs of forensic and anthropological interest using SNaPshot minisequencing, *Forensic Sci Int*, 140, 251-7.

Rosenberg, N.A., Li, L.M., Ward, R. & Pritchard, J.K. 2003. Informativeness of Genetic Markers for Inference of Ancestry, *Am J Hum Genet*, 73, 1402-22.

Salas, A., Bandelt, H.J., Macaulay, V. & Richards, M.B. 2007. Phylogeographic investigations: the role of trees in forensic genetics, *Forensic Sci Int*, 168, 1-13.

Sanchez, J.J., Phillips, C., Børsting, C., Bogus, M., Carracedo, A., Syndercombe-Court, D., Fondevila, M., Harrison, C.D., Morling, N., Balogh, K., et al. 2006. Development of a multiplex PCR assay detecting 52 autosomal SNPs, *International Congress Series*, 1288, 67-9.

Santos, C., Phillips, C., Fondevila, M., Daniel, R., van Oorschot, R.A.H., Burchard, E.G., Schanfield, M.S., Souto, L., Uacyisrael, J., Via, M., et al. 2016. Pacifiplex: an ancestry-informative SNP panel centred on Australia and the Pacific region, *Forensic Sci Int Genet*, 20, 71-80.

Schubert, M., Ginolhac, A., Lindgreen, S., Thompson, J.F., Al-Rasheid, K.A.S., Willerslev, E., Krogh, A. & Orlando, L. 2012. Improving ancient DNA read mapping against modern reference genomes, *BMC genomics*, 13, 178.

Schwark, T., Heinrich, A., Preusse-Prange, A. & von Wurmb-Schwark, N. 2011. Reliable genetic identification of burnt human remains, *Forensic Sci Int Genet*, 5, 393-9.

Scudder, N., McNevin, D., Kelty, S.F., Walsh, S.J. & Robertson, J. 2018a. Forensic DNA phenotyping: Developing a model privacy impact assessment, *Forensic Sci Int Genet*, 34, 222-30.

Scudder, N., McNevin, D., Kelty, S.F., Walsh, S.J. & Robertson, J. 2018b. Massively parallel sequencing and the emergence of forensic genomics: Defining the policy and legal issues for law enforcement, *Science & Justice*, 58, 153-8.

Shafirovich, V. & Geacintov, N.E. 2010, 'Role of Free Radical Reactions in the Formation of DNA Damage', in *The Chemical Biology of DNA Damage*.

Shekar, S.N., Duffy, D.L., Frudakis, T., Sturm, R.A., Zhao, Z.Z., Montgomery, G.W. & Martin, N.G. 2008. Linkage and Association Analysis of Spectrophotometrically Quantified Hair Color in Australian Adolescents: the Effect of OCA2 and HERC2, *Journal Invest Dermatol*, 128, 2807-14.

Shih, S.Y., Bose, N., Gonçalves, A.B.R., Erlich, H.A. & Calloway, C.D. 2018. Applications of Probe Capture Enrichment Next Generation Sequencing for Whole Mitochondrial Genome and 426 Nuclear SNPs for Forensically Challenging Samples, *Genes*, 9, 49.

Sikorsky, J.A., Primerano, D.A., Fenger, T.W. & Denvir, J. 2007. DNA damage reduces Taq DNA polymerase fidelity and PCR amplification efficiency, *Biochem Biophys Res Commun*, 355, 431-7.

Soubrier, J., Gower, G., Chen, K., Richards, S.M., Llamas, B., Mitchell, K.J., Ho, S.Y.W., Kosintsev, P., Lee, M.S.Y., Baryshnikov, G., et al. 2016. Early cave art and ancient DNA record the origin of European bison, *Nat Commun*, 7, 13158.

Sturm, R.A., Duffy, D.L., Zhao, Z.Z., Leite, F.P., Stark, M.S., Hayward, N.K., Martin, N.G. & Montgomery, G.W. 2008. A single SNP in an evolutionary conserved region within intron 86 of the HERC2 gene determines human blue-brown eye color, *Am J Hum Genet*, 82, 424-31.

Sturm, R.A., Teasdale, R.D. & Box, N.F. 2001. Human pigmentation genes: identification, structure and consequences of polymorphic variation, *Gene*, 277, 49-62.

Sulem, P., Gudbjartsson, D.F., Stacey, S.N., Helgason, A., Rafnar, T., Jakobsdottir, M., Steinberg, S., Gudjonsson, S.A., Palsson, A., Thorleifsson, G., et al. 2008. Two newly identified genetic determinants of pigmentation in Europeans, *Nature Genet*, 40, 835.

Taylor, D., Bright, J., McGovern, C., Neville, S. & Grover, D. 2017. Allele frequency database for GlobalFiler™ STR loci in Australian and New Zealand populations, *Forensic Sci Int Genet*, 28, e38-e40.

Templeton, J.E.L., Brotherton, P.M., Llamas, B., Soubrier, J., Haak, W., Cooper, A. & Austin, J.J. 2013. DNA capture and next-generation sequencing can recover whole mitochondrial genomes from highly degraded samples for human identification, *Investig Genet*, 4, 26.

The 1000 Genomes Project Consortium. 2015. A global reference for human genetic variation, *Nature*, 526, 68.

Tofanelli, S., Bertoncini, S., Castrì, L., Luiselli, D., Calafell, F., Donati, G. & Paoli, G. 2009. On the Origins and Admixture of Malagasy: New Evidence from High-Resolution Analyses of Paternal and Maternal Lineages, *Mol Biol Evol*, 26, 2109-24.

Tsai, Y.L. & Olson, B.H. 1992. Rapid method for separation of bacterial DNA from humic substances in sediments for polymerase chain reaction, *Appl Environ Microb*, 58, 2292-5.

United Nations Office of Drugs and Crime 2014, *Global Study on Homicide 2014*, United Nations publication, Sales No. 14.IV.1.

Vallone, P.M. 2012. Capillary electrophoresis of an 11-plex mtDNA coding region SNP single base extension assay for discrimination of the most common Caucasian HV1/HV2 mitotype, *Methods Mol Biol*, 830, 159-67.

Valverde, L., Köhnemann, S., Cardoso, S., Pfeiffer, H. & de Pancorbo Marian, M. 2013. Improving the analysis of Y-SNP haplogroups by a single highly informative 16 SNP multiplex PCR-minisequencing assay, *Electrophoresis*, 34, 605-12.

van Oven, M., Ralf, A. & Kayser, M. 2011a. An efficient multiplex genotyping approach for detecting the major worldwide human Y-chromosome haplogroups, *Int J Legal Med*, 125, 879-85.

van Oven, M., Toscani, K., Tempel, N., Ralf, A. & Kayser, M. 2013. Multiplex genotyping assays for fine-resolution subtyping of the major human Y-chromosome haplogroups E, G, I, J, and R in anthropological, genealogical, and forensic investigations, *Electrophoresis*, 34, 3029-38.

van Oven, M., Vermeulen, M. & Kayser, M. 2011b. Multiplex genotyping system for efficient inference of matrilineal genetic ancestry with continental resolution, *Invest Genet*, 2.

Walsh, S., Liu, F., Wollstein, A., Kovatsi, L., Ralf, A., Kosiniak-Kamysz, A., Branicki, W. & Kayser, M. 2013. The HIrisPlex system for simultaneous prediction of hair and eye colour from DNA, *Forensic Sci Int Genet*, 7, 98-115.

Wang, D.Y., Chang, C.W., Lagace, R.E., Calandro, L.M. & Hennessy, L.K. 2012. Developmental validation of the AmpFlSTR® Identifiler® Plus PCR Amplification

Kit: an established multiplex assay with improved performance, *J Forensic Sci*, 57, 453-65.

Ward, J. 2018. The past, present and future state of missing persons investigations in Australia, *Aust J Forensic Sci*, 50, 708-22.

Wright, K., Mundorff, A., Chaseling, J., Maguire, C. & Crane, D.I. 2015. An Evaluation of the Thai Tsunami Victim Identification DNA Operation, *Forensic Sci Policy Manage*, 6, 69-78.

Xavier, C. & Parson, W. 2017. Evaluation of the Illumina ForenSeq™ DNA Signature Prep Kit – MPS forensic application for the MiSeq FGx™ benchtop sequencer, *Forensic Sci Int Genet*, 28, 188-94.

Y Chromosome Consortium. 2002. A nomenclature system for the tree of human Y-chromosomal binary haplogroups, *Genome Res*, 12, 339-48.

Zar, M.S., Shahid, A.A., Shahzad, M.S., Shin, K.J., Lee, H.Y., Lee, S.S., Israr, M., Wiegand, P. & Kulstein, G. 2018. Forensic SNP Genotyping with SNaPshot: Development of a Novel In-house SBE Multiplex SNP Assay, *J Forensic Sci*.

Zbieć-Piekarska, R., Spólnicka, M., Kupiec, T., Parys-Proszek, A., Makowska, Ż., Pałeczka, A., Kucharczyk, K., Płoski, R. & Branicki, W. 2015. Development of a forensically useful age prediction method based on DNA methylation analysis, *Forensic Sci Int Genet*, 17, 173-9.

Zietkiewicz, E., Witt, M., Daca, P., Zebracka-Gala, J., Goniewicz, M., Jarzab, B. & Witt, M. 2012. Current genetic methodologies in the identification of disaster victims and in forensic analysis, *J Appl Genet*, 53, 41-60.

Zubakov, D., Liu, F., Kokmeijer, I., Choi, Y., van Meurs, J.B.J., van Ijcken, W.F.J., Uitterlinden, A.G., Hofman, A., Broer, L., van Duijn, C.M., et al. 2016. Human age estimation from blood using mRNA, DNA methylation, DNA rearrangement, and telomere length, *Forensic Sci Int Genet*, 24, 33-43.

Chapter 2

A mini-multiplex SNaPshot assay for the triage of degraded human DNA

Manuscript published in *Forensic Science International: Genetics*

Bardan, F., Higgins, D. & Austin, J.J. 2018. A mini-multiplex SNaPshot assay for the triage of degraded human DNA, *Forensic Science International: Genetics*, 34, 62-70.

Statement of Authorship

Title of Paper	A mini-multiplex SNaPshot assay for the triage of degraded human DNA
Publication Status	☑ Published ☐ Accepted for Publication ☐ Submitted for Publication ☐ Unpublished and Unsubmitted work written in manuscript style
Publication Details	Manuscript published in *Forensic Science International: Genetics*. **Bardan, F.**, Higgins, D. & Austin, J.J. 2018. A mini-multiplex SNaPshot assay for the triage of degraded human DNA, *Forensic Science International: Genetics*, 34, 62-70.

Principal Author

Name of Principal Author (Candidate)	Felicia Bardan
Contribution to the Paper	Helped conceive the study, collected and analysed the data, wrote the manuscript and produced the figures.
Overall percentage (%)	70%
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.
Signature	Date 19/10/18

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

i. the candidate's stated contribution to the publication is accurate (as detailed above);

ii. permission is granted for the candidate in include the publication in the thesis; and

iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author	Denice Higgins
Contribution to the Paper	Helped conceive the study, helped collect the samples and assisted in revising manuscript
Signature	Date 20/10/18

Name of Co-Author	Jeremy J Austin
Contribution to the Paper	Helped conceive the study and assisted in revising manuscript
Signature	Date 18 Oct 2018.

Forensic Science International: Genetics 34 (2018) 62–70

Contents lists available at ScienceDirect

Forensic Science International: Genetics

journal homepage: www.elsevier.com/locate/fsigen

Research paper

A mini-multiplex SNaPshot assay for the triage of degraded human DNA

Felicia Bardan*, Denice Higgins, Jeremy J. Austin

Australian Centre for Ancient DNA, School of Biological Sciences, University of Adelaide, North Terrace, Adelaide, South Australia, 5005, Australia

ARTICLE INFO

Keywords:
Degraded DNA
Screening tool
SNPs
Biogeographic ancestry
Sex prediction
Lineage markers

ABSTRACT

Short Tandem Repeat (STR) genotyping is currently the primary DNA-based method for human identification, however it can have limited success when applied to degraded human remains. Massively parallel sequencing (MPS) provides new opportunities to obtain genetic data for hundreds of loci in a single assay with higher success from degraded samples. However, due to the extra requirement for specialised equipment, expertise and resources, routine use of MPS may not be feasible or necessary for many forensic cases. Here we describe the development of a mini-multiplex SNaPshot screening tool (Miniplex) for human samples which allows the qualitative comparison of short mitochondrial and nuclear DNA targets, as well as the interrogation of biogeographic ancestry, lineage, and phenotype single nucleotide polymorphisms (SNPs). This tool is useful to triage samples based on sample quality prior to downstream identification workflows and provides broad biological profile data for intelligence purposes.

1. Introduction

Currently, (STR) profiling is the most widely used method for DNA-based forensic human identification. However, it provides little value to an investigation when there is no match to existing databases or when no STR profile can be obtained due to DNA degradation. In these cases, new methods that interrogate SNPs in mitochondrial, Y chromosome, and autosomal DNA to deliver maternal and paternal lineage, biogeographic ancestry, and/or phenotypic information can provide intelligence to guide identification attempts [1]. This information can direct targeted collection of reference samples in cases where obtaining a STR profile is likely, or can be used as the primary DNA evidence in cases where no STR profile can be obtained. A number of methods and panels have been developed for use in human identification, most of which focus on a single biological query, such as mtDNA [2,3], Y chromosome [4], biogeographic ancestry [5,6], identity SNPs [7] or phenotype [8]. Hence to establish a comprehensive understanding of a person's biological profile, multiple sequential tests must be performed with significant cost and time burdens as well as consumption of valuable sample.

Recent developments in MPS of polymerase chain reaction (PCR) multiplexes offer a solution to this issue, via simultaneous sequence analysis of hundreds of genetic markers [9,10]. However this technology has a number of limitations, including cost and time constraints, and the need for additional specialised equipment and expertise. These can be significant hurdles for routine adoption in forensic biology laboratories, especially those with low-throughput MPS requirements. Commercial MPS kits have also shown limitations with performance when analysing degraded DNA, both in the recovery of longer genetic targets (e.g. the maximum amplicon length for the ForenSeq™ panel is 430 bp) [11,12] and in the interpretation of data [13].

Constraints imposed by limitations on available resources, expertise and biological sample mean that careful decisions must be made to determine which samples should be analysed with more specialised methods. A screening tool to assess the presence and quality of both nuclear and mitochondrial DNA in a sample and to provide broad profiling information about the donor (maternal and paternal lineage, biogeographic ancestry and phenotype) would be beneficial to triage samples to allow appropriate use of available resources and to assist in directing an investigation.

Here we describe the development of a multi-target SNP-based screening panel ('Miniplex'). This new panel interrogates 18 markers that are informative for mtDNA and Y chromosome (Y-chr) haplogroups, biogeographic ancestry, and phenotype (eye colour). The Miniplex is a quick, efficient and economical method with simple data analysis that delivers a broad biological profile in a single multiplex assay. Data generated can be used as a preliminary screening tool to inform subsequent analyses by aiding in directing resources and management of samples (i.e. to determine which samples require more specialised methods). The Miniplex provides a measure of DNA quality via qualitative comparison of mtDNA and nuclear SNP recovery, and DNA availability in degraded samples by targeting short amplicons. This panel also allows sex determination due to the inclusion of multiple Y-chr SNPs. The benefits of combining different marker types are

* Corresponding author.
E-mail address: felicia.bardan@adelaide.edu.au (F. Bardan).

https://doi.org/10.1016/j.fsigen.2018.02.006
Received 15 September 2017; Received in revised form 3 January 2018; Accepted 5 February 2018
Available online 06 February 2018

especially applicable for populations where admixture is a factor in their demographic history.

2. Materials and methods

2.1. SNP selection and multiplex design

We selected 17 SNPs and one indel from previously published panels [2–6,8]; to allow broad inference of mtDNA haplogroup, Y-haplogroup, biogeographic ancestry (mtDNA, Y −chr and autosomal) and phenotype (eye colour). These include five SNPs to define global mtDNA haplogroups (L3*, M*, N*, R and D), four SNPs and one indel for global Y- chr haplogroups (D, E, C, R and O), five SNPs that differentiate between five continental biogeographic ancestry groups (Africa, Europe, East Asia, Oceania and Native America), and three SNPs for eye colour prediction (Supplementary Table S1).

PCR and single base extension (SBE) primer sequences were obtained from the original publications (Supplementary Tables S1 and S2) with the exception of PCR primers for SNP rs12913832 (eye colour) where new PCR primers were designed to reduce amplicon size, following guidelines set out in [14]. The length of PCR amplicon sizes range from 66 to 128 bp. All primers were screened for primer-dimer formation, complementarity and hairpin interaction using Primer-BLAST [15] and Multiplex Manager [16]. SBE primers were high performance liquid chromatography (HPLC) purified and modified at the 5′ end with poly-CT tails to ensure appropriate fragment length spacing within the four fluorescent dye channels. The length of the SBE primers was between 24 and 90 bp, and primers using identical dye combinations were separated by at least three bp in length.

2.2. Reference samples and DNA extractions

We used high quality reference DNA samples to test and optimise the multiplex. Buccal swabs on FTA cards were obtained from human donors with informed consent in accordance with ethics approval from the University of Adelaide Human Research and Ethics Committee (H-2016-218). Donors had a range of self-declared biogeographic ancestry (Europe, Africa, Native America, East Asia) and included people with blue, intermediate and brown eyes. DNA was extracted from ~5 mm^2 pieces of FTA card using a Charge Switch Forensic DNA Purification Kit (Thermo Fischer Scientific) following manufacturer's instructions and quantified using the Qubit (Life Technologies) high sensitivity assay (concentrations given in Supplementary Table S3).

2.3. Multiplex PCR, SNaPshot and capillary electrophoresis (CE) conditions

To assess the performance of each primer pair and SBE probe, to confirm the SNP at each marker, and to set-up the custom panel and bin settings, all PCR primer pairs and SBE probes were tested in singleplex on male and female reference DNA samples. PCR, SBE and CE conditions for the singpleplexes were identical to the multiplex protocol described below except the final primer concentrations were 0.8 uM for PCR primers and 160 nM for SBE probes. For bins unable to be resolved using available reference DNA samples, we designed three 105 bp oligonucleotides containing the SNP of interest for use as template (8 nM final concentration per PCR) in singleplex PCR and SBE reactions. We attempted to balance the final multiplex SNaPshot profile peak heights by adjusting the PCR and SBE primer concentrations based on the relative fluorescence units (rfu) of the initial singleplexes, with subsequent adjustment in the multiplex.

The optimised multiplex PCR protocol used 12.5 uL volumes containing 1uL of DNA extract, 1.25 uL of 10× ImmoBuffer, 1.125 uL of 50 mM $MgCl_2$, 0.1 uL Immolase DNA Polymerase (Bioline Pty Ltd), 1 uL of 10 mM dNTP mix (Applied Biosystems), 0.2 uL of 3.2 mg/ml RSA (Sigma- Aldrich), 5.55 uL H_20 and 2 uL of primer mix consisting of 18 primer pairs with final concentrations given in Supplementary Table S1. PCRs were carried out on a T1000 Thermal Cycler (Bio-Rad Laboratories) using the following conditions: 95 °C for 10 min followed by 30 cycles of 94 °C for 45 s, 60 °C for 45 s, 72 °C for 60 s and a final extension at 72 °C for 6 min. Amplification success was assessed by gel electrophoresis on a 3.5% agarose gel (100 V for 45 min; Hyperladder V DNA size ladder (Bioline Pty Ltd)). PCR products were purified by combining 2.5 uL PCR product with 1 uL Illustra ExoProStar PCR cleanup reagent (GE Healthcare Life Sciences), followed by incubation at 37 °C for 45 min and 80 °C for 15 min.

SBE reactions consisted of a final volume of 3 uL containing 1 uL purified PCR product, 1.25 uL SNaPshot® Ready Reaction mix (Applied Biosystems), and 0.25 uL of a SBE primer mix with final concentrations detailed in Supplementary Table S2. Thermocycling was performed on a T1000 Thermal Cycler (Bio-Rad Laboratories) with the following conditions: 96 °C for 2 min followed by 30 cycles of 96 °C for 10 s, 55 °C for 5 s, 60 °C for 30 s. SBE products were purified by adding 1 uL of Illustra Alkaline Phosphatase (GE Healthcare Life Sciences) and incubating for 37 °C for 80 min followed by enzyme inactivation at 85 °C for 15 min.

One uL of purified SBE product was mixed with 9.75 uL Hi-Di Formamide and 0.25 uL GeneScan-120 LIZ internal size standard (Applied Biosystems). Capillary electrophoresis was performed on a 3500 Genetic Analyzer (Applied Biosystems) with 36 cm arrays and POP-4 polymer using a customised run module. Electropherograms were analysed for genotype calling in GeneMapper ID version 5.0 (Applied Biosystems) using a custom panel and bin settings.

2.4. Multiplex profile interpretation

As the assay is a multi-target SNP genotyping tool, each marker type requires separate interpretation for an overall assessment of mtDNA and Y-chr haplogroup, biogeographic ancestry and phenotype.

2.4.1. Lineage markers

Y-chr and mtDNA SNPs were mapped onto the Y-chr [17] and mtDNA [18] trees on PhyloTree build 17 (www.phylotree.org) to predict the Y-chr and mtDNA haplogroup respectively, and to check for phylogenetic sense (i.e. the SNP profile predicted a single haplogroup).

2.4.2. Autosomal ancestry

The autosomal SNPs were compared to genotypes from 402 individuals across five reference population groups (Africa = AFR, Europe = EUR, East Asia = EAS, America = AMR and Oceania = OCE) from the 1000 Genomes Project [19], and the CEPH human genome diversity panel (HDGP-CEPH) [20] (see Supplementary file S1) using the online Bayesian forensic classifier Snipper (http://mathgene.usc.es/snipper/). AFR, EUR and EAS population genotypes for the five autosomal ancestry SNPs in the Miniplex were sourced from samples of the 1000 Genomes Project, and AMR and OCE genotypes were sourced from the HGDP-CEPH dataset [5]. Likelihood ratios (LR) for the biogeographic ancestry classifications were generated from the output, and a 2D PCA plot generated in R as described in a previous study [21], and were used to inform biogeographic ancestry.

2.4.3. Phenotype

For application as a screening tool, and because of the complex polygenic nature of intermediate iris colour predictors [8], the three phenotype SNPs included in this panel were chosen to provide an indication of 'brown' versus 'not brown' eye colour. Phenotype SNPs were analysed to predict eye colour (brown or not brown) using the prediction model from the HIrisPlex Eye and Hair Colour DNA Phenotyping Webtool (hirisplex.erasmusmc.nl) outlined in a previous study [8].

2.4.4. Sex

The Y-chr SNPs were used to identify the sex of each DNA donor.

The presence of two or more Y SNPs was interpreted as a male. However, due to the possibility of degraded DNA, the absence of Y-chr SNPs was only used to indicate a female if mtDNA, ancestry and phenotype SNPs were obtained.

2.5. Quality control, concordance, and sensitivity tests

We tested the Miniplex on ten reference DNA samples (Section 2.2), comprising five males and five females with known mtDNA haplogroup, self-declared ancestry and phenotype. The mtDNA haplogroups were obtained via Sanger Sequencing of the control region with haplogroup prediction using EMPOP (https://empop.online/).

2.5.1. Quality controls

We included a positive control, of known sex and genotype, in all PCR and SBE typing attempts to ensure reproducibility between batches. Negative extraction controls and PCR negative controls were also included to monitor for contamination, allele and locus drop-in and other artefacts.

2.5.2. Concordance

The SNP profile obtained from the Miniplex assay for one reference sample was compared to sequence data obtained via MPS amplicon sequencing. Multiplex PCR was carried out as described in 2.3. Library preparation was undertaken using the KAPA Hyper Prep Kit (KAPA Biosystems) following the manufacturer's instructions. In brief, PCR product was end repaired & A-tailed on 3′ ends. Barcoded library adapters were then ligated and purified with Ampure (Agencourt). Library amplification was performed using seven cycles on a T1000 Thermal Cycler (Bio-Rad Laboratories), followed by Ampure (Agencourt) purification. The library was quantified using the Agilent 2200 TapeStation (Agilent Technologies). Sequencing was performed on an Illumina MiSeq at Australian Genome Research Facility using paired end 150 bp reads.

2.5.3. Sensitivity testing

To test the sensitivity threshold of the assay, a male and female control DNA sample (Promega) were genotyped in a doubling dilution series of DNA with input levels of 1 ng, 0.5 ng, 0.25 ng, 125 pg, 64 pg, 32 pg, and 16 pg. Each dilution was run in triplicate and negative and positive controls were included in all PCR and SBE runs.

2.6. Testing on degraded DNA samples

We subsequently tested the Miniplex on degraded DNA samples using extracts from degraded human teeth [22] in accordance with ethics approval from the University of Adelaide Human Research and Ethics Committee (H-2016-198). Known ancestry and phenotype data were not available for these samples. In brief, isolated human teeth were buried in soil for periods of time ranging from one to 16 months, and then DNA was extracted from cementum tissue as described in the study [22]. All DNA extracts were quantified using the Qubit (Life Technologies) high sensitivity assay (sample information and DNA concentrations given in Supplementary Table S4). STR typing success for each sample was also assessed and data was made available from the previous study [22]. Three teeth from zero months acted as quality controls, and four teeth from each post-mortem interval (1, 2, 4, 8, and 16 months) were used.

3. Results

3.1. The miniplex assay: design and sensitivity

Despite efforts to achieve peak height balance, mtDNA SNPs consistently displayed higher rfu values than nuclear SNPs (Fig. 1). The panel performed optimally at DNA input amounts between 1 ng and 64 pg, however full profiles were obtained with 500 pg, 250 pg, 125 pg, 64 pg, and 32 pg of input DNA. For 16 pg of input DNA, all mtDNA SNPs were recovered, and locus and allele dropout occurred for nuclear DNA SNPs, with no less than 85% recovery rate.

3.2. Performance on reference DNA samples

3.2.1. mtDNA and Y-chr SNPs

We obtained all five mtDNA SNPs from the ten reference DNA samples, and all five Y-chr SNPs from the five male reference DNA samples. All five female samples displayed a peak for Y-chr locus M174. All mtDNA and Y-chr SNPs made phylogenetic sense (i.e. they predicted only a single haplogroup), and the inferred haplogroups were concordant with self-declared ancestry (Table 1). The mtDNA haplogroup inferred from the five SNP profile was concordant with the EMPOP predicted haplogroup from the control region sequence.

3.2.2. Biogeographic ancestry

We obtained all five biogeographic ancestry SNPs from all ten reference DNA samples. Biogeographic ancestry predictions from Snipper were concordant with self-declared ancestry (Table 2).

The PCA plot (Fig. 2) based on five autosomal ancestry SNPs for the ten reference samples and 402 population sample genotypes shows clear separation of African and European samples, but weak and/or overlapping distribution for American, Oceanian, and East Asian samples. The position of the ten reference DNA samples is broadly consistent with their self-declared biogeographic ancestry.

3.2.3. Phenotype prediction

We obtained all three phenotype SNPs from each of the ten reference DNA samples. The HIrisPlex Eye and Hair Colour DNA Phenotyping Webtool correctly predicted eye colour as 'brown' or 'not brown' (Table 3). As the full suite of HIrisplex SNPs were not typed in this assay, the loss in 'area under curve' (AUC), which represents the loss in accuracy of the prediction, is reported due to missing data. Our results show that this panel is still able to indicate brown and not brown eye colour with a high probability and low AUC loss.

3.2.4. Sex prediction

All five male samples displayed peaks for all five Y-chr SNPS. As mentioned above (Section 3.2.1) all five females displayed a peak for the ancestral allele T of Y-chr SNP M174. The sex of all ten reference DNA donors was therefore predicted accurately on the basis of the presence or absence of two or more Y-chr SNP markers.

3.3. Concordance study

3.3.1. Singleplex sanger sequencing

Sanger sequencing of mitochondrial control region and EMPOP prediction for each sample was concordant with the five SNP profile haplogroup prediction obtained from the 'Miniplex' panel.

3.3.2. MPS amplicon sequencing

All 18 markers were concordant with MPS amplicon sequencing results (data not shown). The read depth per locus ranged from 2843 reads for Y-chr SNP M216 to 15,752 reads for mtDNA SNP 12705, with a mean of 7305 reads per SNP.

3.4. Application to low quality degraded DNA

3.4.1. Profiling success

We retrieved all five mtDNA SNPs from all 23 degraded tooth samples, and full nuclear SNP profiles (13 SNPs for males, 8 SNPs for females) from nine out of the 23 samples (Fig. 3). Nuclear SNP typing success was equal to or higher than STR typing success in all samples. For the 14 samples displaying partial profiles, nuclear SNP success

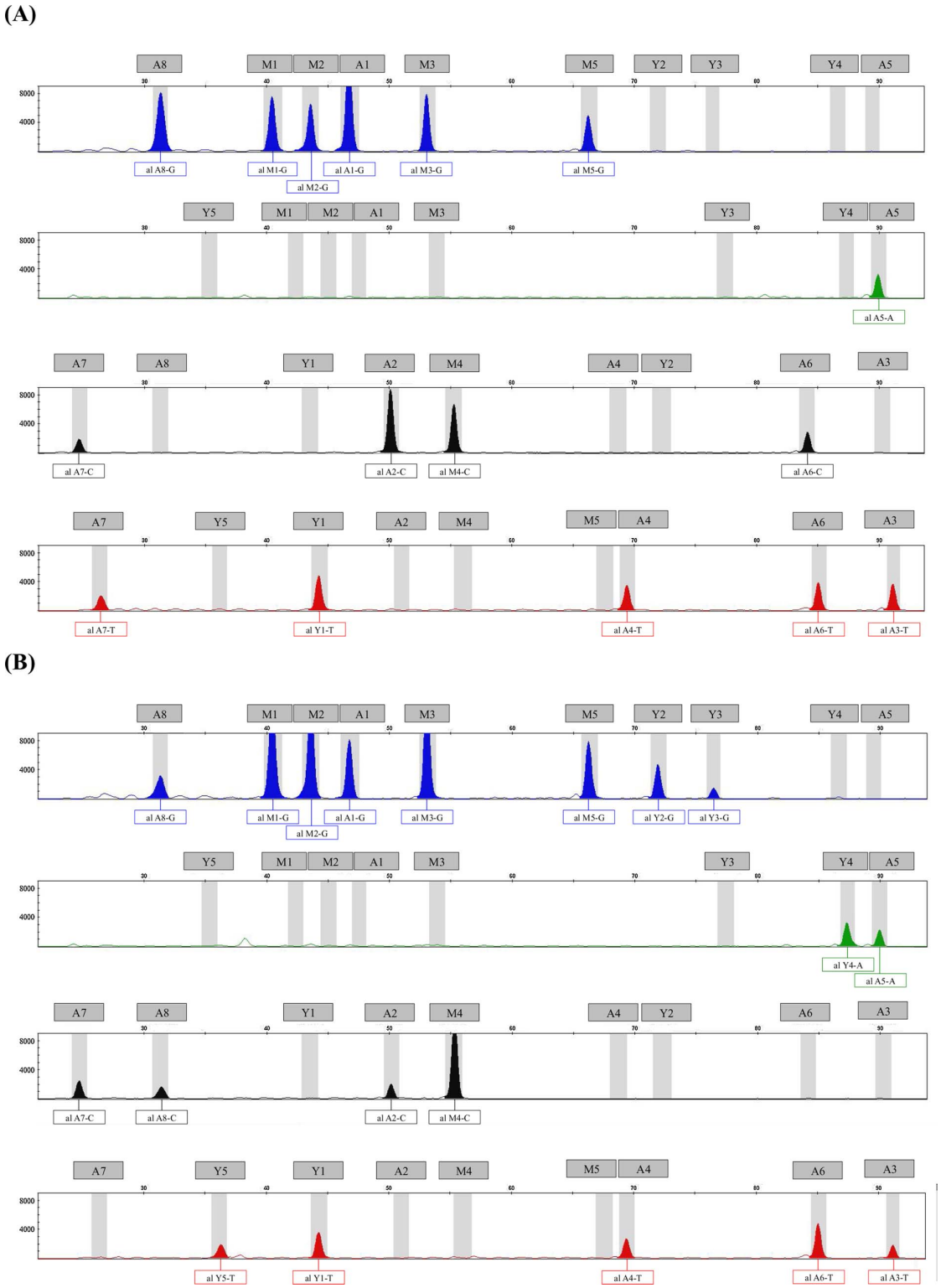


Fig. 1. Example electropherograms for the Miniplex for (A) female, and (B) male reference samples with 1 ng of input DNA. Vertical grey boxes represent allele bins. Alphanumeric codes above grey boxes refer to the locus code: M1-M5 = mtDNA, Y1-Y5 = Y chr, A1–A8 = autosomal ancestry and phenotype (see Supplementary Table 1 for details).

Table 1
Sample information for ten reference DNA samples, five males and five females, and their inferred mtDNA and Y-chr haplogroup from the Miniplex compared to self-declared ancestry and mtDNA haplogroup from control region sequence. *Indicates mtDNA macro-haplogroup.

Sample	Sex	Self-declared ancestry	mtDNA haplogroup (control region sequence)	Inferred mtDNA haplogroup	Inferred Y chr haplogroup
ACAD5	Male	East Asian	M17c1a1	M*	O
ACAD17	Male	Native American	A6b	N*	R
ACAD34	Male	European	U2e1	N*	R
ACAD22	Male	African	L3e3b	L3*	E
ACAD35	Male	East Asian	Z3	M*	O
ACAD8	Female	European	T2b	R	na
ACAD1	Female	European	J2b1	R	na
ACAD32	Female	European	T2f1a	R	na
ACAD33	Female	East Asian	B4c1c	R	na
ACAD29	Female	European	J1c2	R	na

ranged from 12.5–92%, compared to 0–90% STR success. The 'Miniplex' generated partial nuclear SNP profiles and full mtDNA profiles from three samples that produced no STR data previously, and an increased percentage of target recovery compared to previous STR data for 16 samples.

3.4.2. mtDNA and Y SNPs

We obtained all five mtDNA SNPs from all samples, and five of eleven male samples produced all five Y-chr SNPs (Table 4). Y-chr SNP dropout occurred most frequently for the M175 indel. The mtDNA and Y-chr SNP profiles made phylogenetic sense and indicated only one haplogroup each. Three male samples were not assigned to any of the haplogroups predicted by the Miniplex and were labelled as 'Not D, E, C, R, O'. Four male samples had insufficient Y-chr SNPs typed to assign a Y-haplogroup.

3.4.3. Biogeographic ancestry

Twenty samples returned a 'European' classification, and one sample returned a 'Native American' classification using Snipper (Table 5). Two samples could not be classified due to missing data.

The PCA plot (Fig. 4) for the five ancestry SNPs using 402 reference population genotypes and degraded teeth sample shows a clustering of the samples around the European reference population, with sample 4102 falling around the East Asian, Oceanian and American reference populations.

3.4.4. Phenotype

Seventeen out of 23 samples produced sufficient SNPs to generate an eye colour prediction using the HIrisplex Hair Colour DNA Phenotyping Webtool. Thirteen of these samples produced all three phenotype SNPs. Eleven samples for which a prediction was made were predicted as having 'not brown' eye colour and six samples were predicted as a 'brown' eye colour phenotype (Table 6). Six of the 23 samples had missing data for which the webtool was not able to make a prediction.

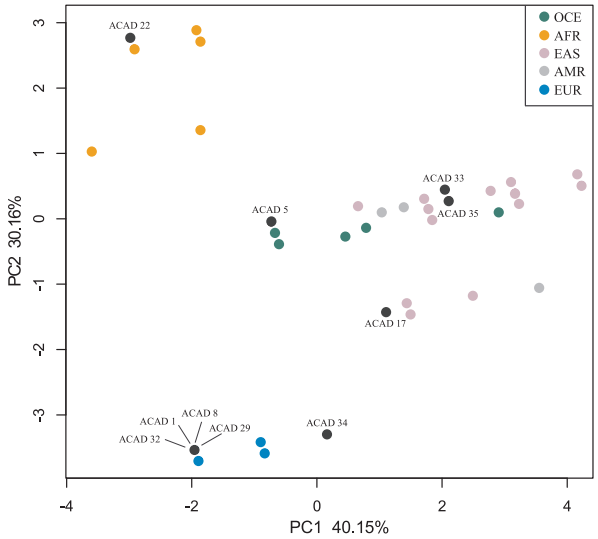


Fig. 2. PCA plot of ten reference samples with known (self-declared) ancestry (black circles) and 402 population genotypes representing AFR (orange), EUR (blue), AMR (grey), EAS (pink) and OCE (green) populations from the 1000 Genomes or HGDP-CEPH datasets. Single coloured dots may represent > 1 population reference sample where multiple individuals shared an identical five locus SNP genotype. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 3
Inferred eye colour, and their associated probabilities and AUC loss due to missing data (i.e. SNPs not typed from the HIrisplex assay) for ten reference samples with known eye colour using the Miniplex assay.

Sample	Reported Eye Colour	Inferred Eye Colour	p-value	AUC loss
ACAD5	Brown	Brown	0.996	0.013
ACAD17	Brown	Brown	0.974	0.013
ACAD34	Brown	Brown	0.99	0.013
ACAD22	Brown	Brown	0.996	0.013
ACAD35	Brown	Brown	0.96	0.013
ACAD8	Blue	Not Brown	0.973	0.018
ACAD1	Intermediate	Not Brown	0.559	0.065
ACAD32	Intermediate	Not Brown	0.973	0.018
ACAD33	Brown	Brown	0.996	0.013
ACAD29	Blue	Not Brown	0.973	0.018

Table 2
Inferred biogeographic ancestry for ten reference samples with self-declared ancestry using the Miniplex and Snipper, and their associated likelihood ratios. The lowest and highest likelihood ratios are presented to demonstrate the range of values obtained. Remaining likelihood ratios are given in Supplementary Table S5).

Sample	Self-declared ancestry	Snipper inferred ancestry	Lowest and Highest Likelihood Ratio from Snipper
ACAD5	East Asian	East Asian	9 times more likely EAS than OCE and 86.49 times more likely EAS than EUR
ACAD17	Native American	Native American	1.14 times more likely AMR than EAS and 5,037,195 times more likely AMR than AFR
ACAD34	European	European	531 times more likely EUR than AMR and 790,716,457 times more likely EUR than AFR
ACAD22	African	African	15,073,563 times more likely AFR than OCE, and 1,772,315,995 times more likely AFR than EUR
ACAD35	East Asian	East Asian	874 times more likely EAS than AMR and 2,533,050,134 times more likely EAS than AFR
ACAD8	European	European	514,910 times more likely EUR than EAS, and 13,071,541 times more likely EUR than OCE
ACAD1	European	European	514,910 times more likely EUR than EAS, and 13,071,541 times more likely EUR than OCE
ACAD32	European	European	514,910 times more likely EUR than EAS, and 13,071,541 times more likely EUR than OCE
ACAD33	East Asian	East Asian	52.21 times more likely EAS than AMR, and 889,526,628 times more likely EAS than AFR
ACAD29	European	European	514,910 times more likely EUR than EAS, and 13,071,541 times more likely EUR than OCE

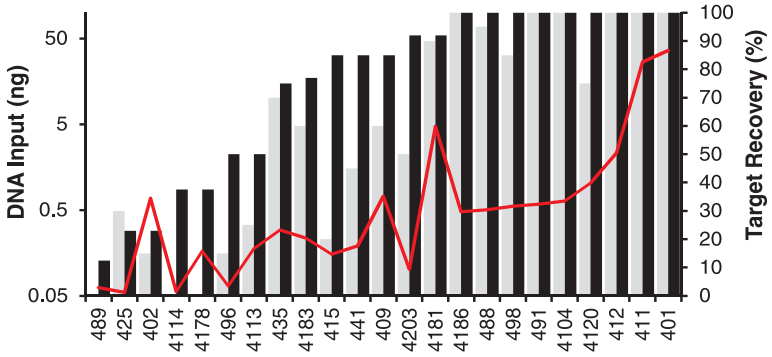


Fig. 3. Miniplex nuclear SNP typing success (black bars) on a range of degraded human teeth samples with varying DNA input amounts (red line) and where previous STR success (grey bars) is known. MtDNA SNP success not shown due to all samples returning complete mtDNA SNP profiles. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 4
Lineage marker results from degraded DNA from 23 buried human teeth samples using five mtDNA and five Y-chr SNPs in the Miniplex. 'na' denotes samples for which a Y-chr haplogroup was not assigned. 'Not D, E, C, R, O' denotes samples which did not fall into any of the Miniplex Y haplogroups.

Sample	Sex	Inferred mtDNA haplogroup	Inferred Y-chr haplogroup
401	Male	R	R
411	Female	R	na
412	Female	R	na
491	Female	R	na
4183	Male	N	na
4104	Male	R	R
4113	Female	R	na
4114	Female	R	na
4186	Female	R	na
4203	Male	M	na
4181	Male	R	Not D, E, C, R, O
441	Male	R	Not D, E, C, R, O
4178	Female	R	na
4120	Male	M	Not D, E, C, R, O
435	Female	R	na
409	Female	R	na
498	Male	R	R
402	Male	R	na
488	Female	R	na
415	Male	R	R
489	Female	R	na
496	Female	R	na
425	Male	M	na

3.4.5. Sex

We obtained all five Y-chr markers from five of the 11 known male samples. Nine male samples produced between two and five Y-chr markers and so were correctly predicted as male. Two male samples could not be predicted correctly due to obtaining only one Y-chr marker. These samples exhibited poor success with previous STR typing, and had limited success with other SNPs typed in the Miniplex. According to our criteria for sex prediction, all female samples were correctly predicted. Female samples still displayed a peak for Y marker M174.

4. Discussion

We have developed and tested a SNaPshot mini-multiplex assay that provides a screening tool to help direct forensic investigations of unknown human DNA samples. The Miniplex targets multiple mtDNA, Y-chr and autosomal SNPs and one indel that provide an indication of sample quality (mtDNA vs nuclear DNA, with an average amplicon length of 91 bp); lineage marker haplogroup; biogeographical ancestry (autosomal and maternal/paternal lineage); eye colour and sex to inform and tailor downstream processes for more detailed genetic analysis. The panel is intended as a rapid and cost effective presumptive tool to provide an assessment of DNA quality and a broad biological profile, prior to more detailed, and costly, genetic analyses. Investigations involving degraded DNA, missing persons, war dead and samples with STR profiles that do not match any existing database could benefit from this panel.

For samples at and above the empirically determined sensitivity threshold of 32 pg input DNA, we obtained 100% SNP typing success where predicted biological profiles were consistent with self-declared ancestry and phenotype. Additionally, SNP results were concordant with MPS and Sanger sequencing results. The sensitivity of the Miniplex assay is comparable to other SNaPshot assays used in forensic analysis of degraded DNA where full profiles are able to be obtained from samples down to 64 pg [5] and 31 pg of input DNA [23]. Taken together, these suggest that the Miniplex is both sensitive and accurate, and could provide a useful new tool to triage degraded human samples.

The Miniplex is the only forensic multiplex that combines mtDNA, autosomal DNA, and Y-chr markers in a single assay. As expected, due to the higher copy number of mtDNA, the mitochondrial SNPs in this panel outperformed the nuclear markers, with higher peak heights and 100% recovery for all samples. Attempts were made during optimisation to minimise the peak height imbalance however it will always be a consideration for panels with large differences in target copy number, such as mtDNA and nuclear DNA.

Unlike existing SNaPshot multiplexes that focus on a single biological query (mtDNA, Y-chr, autosomes), the Miniplex provides a qualitative comparison of mtDNA and nuclear DNA preservation in a sample to allow an assessment of DNA availability and degradation. As is well known, the higher copy number and more robust structure of mtDNA over nuclear DNA allows it to survive for longer post-mortem intervals. A full mtDNA SNP profile with a partial nuclear SNP profile indicates more advanced DNA degradation with limited availability of short nuclear DNA fragments and a lower likelihood of STR-profiling success (Fig. 3). This can help to inform downstream processing as to whether mtDNA should be further interrogated as the focus (applicable to the more degraded samples in this study). It may also help to decide whether PCR-based STR typing will be unsuccessful and thus a sample needs to be subjected to more specialised methods for target recovery such as those involving short amplicon sequencing via MPS (e.g. [24]).

Biogeographic ancestry assignment is increasingly important in challenging human identification cases, especially where anthropological information is missing. Recent work has shown that small-scale multiplexes of less than 35 SNPs can provide discrimination between continental-scale ancestral populations and identify admixture [5,6]. Using only five of the most discriminatory SNPs from the Global AIMs Nano set [5], the Miniplex correctly predicted the biogeographic ancestry of samples with known ancestry. However, due to the limited number of SNPs, two individuals originating from recently admixed populations (ACAD5–South-East Asia and ACAD17–Central America) had lower likelihood ratios. Both regions have been shown to contain high levels of recent admixture [25,26] complicating biogeographic ancestry prediction. Despite the low likelihood ratios obtained for some

Table 5
Snipper ancestry predictions and their associated likelihood ratios applied to degraded DNA from 23 buried human teeth, using five biogeographic ancestry SNPs in the Miniplex. The lowest and highest likelihood ratios are presented to demonstrate the range of values obtained. Remaining likelihood ratios are given in Supplementary Table S6).

Sample	Snipper inferred ancestry	Lowest and Highest Likelihood Ratio from Snipper
401	European	514,910 times more likely EUR than EAS, and 202,074,702 times more likely EUR than AFR
411	European	18,085 times more likely EUR than OCE, and 353,615,244 times more likely EUR than AFR
412	European	514,910 times more likely EUR than EAS, and 202,074,702 times more likely EUR than AFR
491	European	514,910 times more likely EUR than EAS, and 202,074,702 times more likely EUR than AFR
4183	European	514,910 times more likely EUR than EAS, and 202,074,702 times more likely EUR than AFR
4104	European	514,910 times more likely EUR than EAS, and 202,074,702 times more likely EUR than AFR
4113	European	3646 times more likely EUR than OCE, and 236,834,750 times more likely EUR than EAS
4114	European	569 times more likely EUR than AMR, and 4, 825 times more likely EUR than AFR
4186	European	514,910 times more likely EUR than EAS, and 202,074,702 times more likely EUR than AFR
4203	European	47,364 times more likely EUR than AMR, and 222,694,403 times more likely EUR than AFR
4181	European	514,910 times more likely EUR than EAS, and 202,074,702 times more likely EUR than AFR
441	European	47,364 times more likely EUR than AMR, and 222,694,403 times more likely EUR than AFR
4178	European	652 times more likely EUR than AMR, and 237,042,214 times more likely EUR than AFR
4120	American	36.94 times more likely AMR than EAS, and 178, 762 times more likely AMR than AFR
435	European	572 times more likely EUR than AMR, and 222, 889, 346 times more likely EUR than AFR
409	European	47,364 times more likely EUR than AMR, and 222,694,403 times more likely EUR than AFR
498	European	514,910 times more likely EUR than EAS, and 202,074,702 times more likely EUR than AFR
402	NA	Could not be classified
488	European	514,910 times more likely EUR than EAS, and 202,074,702 times more likely EUR than AFR
415	European	514,910 times more likely EUR than EAS, and 202,074,702 times more likely EUR than AFR
489	European	569 times more likely EUR than AMR, and 222, 889, 346 times more likely EUR than AFR
496	European	3555 times more likely EUR than OCE, and 53, 688 times more likely EUR than AMR
425	NA	Could not be classified

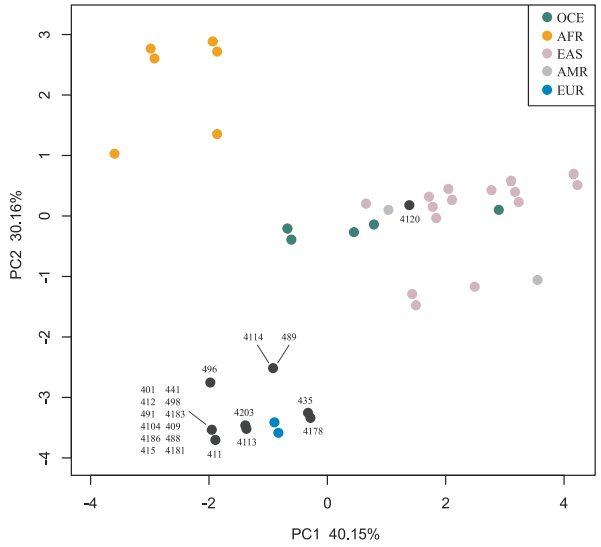


Fig. 4. PCA plot of 21 degraded DNA samples (black circles) with 402 population genotypes representing AFR (orange), EUR (blue), AMR (grey), EAS (pink) and OCE (green) populations from 1000 Genomes or HGDP-CEPH datasets. Single coloured dots may represent > 1 population reference sample where multiple individuals shared an identical five locus SNP genotype. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 6
Inferred eye colour for degraded DNA from 23 buried human teeth, using three phenotype SNPs in the Miniplex and the HIrisplex webtool. 'na' denotes samples for which insufficient SNPs were typed and AUC loss was too large to make a prediction.

Sample	Inferred Eye Colour	p-value	AUC loss
401	Not Brown	0.973	0.018
411	Not Brown	0.907	0.018
412	Brown	0.704	0.013
491	Brown	0.704	0.013
4183	Brown	0.704	0.013
4104	Not Brown	0.973	0.018
4113	na	na	0.416
4114	na	na	0.416
4186	Brown	0.974	0.013
4203	Not Brown	0.973	0.018
4181	Brown	0.704	0.013
441	Not Brown	0.973	0.018
4178	na	na	0.416
4120	Brown	0.996	0.013
435	Not Brown	0.559	0.065
409	Not Brown	0.970	0.083
498	Not Brown	0.973	0.018
402	Not Brown	0.970	0.1
488	Not Brown	0.973	0.018
415	Not Brown	0.970	0.083
489	na	na	na
496	na	na	0.364
425	na	na	na

samples between East Asian, Oceanian and Native American ancestry (ACAD5, ACAD17 and ACAD33), the five ancestry SNPs do provide definitive ancestry exclusion. ACAD5, ACAD17 and ACAD33 are 86, 5 million and 889 million times more likely East Asian, Native American and East Asian, respectively, than African, thus excluding African ancestry for all three individuals. Admixture remains a concern even for SNaPshot panels with a larger number of SNPs, especially for populations that show reduced divergence (hence weaker pairwise differentiation) due to recent shared history [6]. For example, a reduced pairwise differentiation of individuals from East Asia/Oceania, as well as East Asia/Native America has been demonstrated in previous panels [5,6]. The inclusion of mtDNA and Y-chr haplogroup defining SNPs provides one (females) or two (males) independent ancestry predictions that may help to overcome the limitations of using only five autosomal ancestry markers. As an additional consideration, the addition of markers discriminatory for particular regions of interest can be easily implemented for adapting the panel for more specific questions of ancestry. For example, in an Australian context, the inclusion of markers indicative of Australian Aboriginal ancestry (e.g. Y chromosome Hg K and mtDNA Hg P or S) could be beneficial in the Miniplex to direct investigations and screen for remains from Australia, particularly for traditional Aboriginal Australian burials where restrictions on genetic data is of importance.

The Miniplex Y-chr SNPs allowed prediction of sex in samples with sufficient nuclear DNA and in no instance was sex predicted incorrectly.

The presence of a peak for marker M174 in female samples is due to the high level of homology of this locus in the ubiquitin-specific protease 9 (USP9) gene between the X and Y chromosome and has been observed in a previous study [27]. This could not be overcome unless the amplicon was made much longer and therefore would be unsuitable for degraded DNA. We considered replacing this locus with an alternative SNP however this marker was judged as the most appropriate for the indication of Y-haplogroup D. Thus, we ultimately decided to retain this marker and account for the appearance of a peak in female samples in our analysis.

Phenotype prediction with a limited number of SNPs is always going to be challenging, as they are complex genetic traits determined by the cumulative effect of many genes [28]. Hence for this panel we only intended to distinguish between brown and non-brown eye colour. Intermediate eye colour can be difficult to predict even with a much larger number of SNPs [13,29,30], but is also subject to variable interpretation by the observer. Given that the Miniplex is a presumptive tool with a much smaller number of SNPs, this effect can be exacerbated and as such we only aimed to distinguish between brown and not-brown eye colour. Eye colour is less likely to change over a person's lifetime than hair colour, which can be affected by age and chemical treatments, so is more informative for intelligence purposes. Despite only using three eye colour SNPs from the HIrisplex panel, the Miniplex was able to correctly predict the eye colour class (brown or not-brown) for all ten reference samples with a high *p*-value and low AUC loss.

The Miniplex provided genetic information for a range of reference and degraded DNA samples, including those that had limited or no previous success with STR typing. This demonstrates the feasibility of the panel as a tool that could not only be performed prior to STR typing as an indicator of STR success, but as a complementary SNP typing tool to obtain a broadly indicative biological profile, both of which may help to inform downstream processing. Partial profiles were obtained from samples that had previously exhibited poor STR success as anticipated, whereas those which had high success with STR typing produced full profiles, indicating that the Miniplex can be used as a measure of DNA quality. As this panel is intended as a presumptive screening tool, we recommend verifying biological profiles obtained using the Miniplex by confirmatory testing.

5. Conclusion

This panel was designed to be an efficient and sensitive tool to cope with highly degraded and low template DNA, a common concern when analysing forensic samples. The assay also tests for the presence of short endogenous DNA fragments as an indication for which methods will likely be successful in analysing the sample. Measuring the quality and presence of short DNA sequences to triage samples is a useful tool not only in traditional forensic biological testing but more importantly to allocate resources to samples in the present next-generation sequencing era which can be a lengthy and expensive process, particularly for low-throughput MPS laboratories. This panel has demonstrated promise for use in identification investigations and uses conventional capillary electrophoresis technologies that are already validated and optimised for DNA analyses in forensic laboratories. Lastly, we present the only tool developed for triaging forensic samples that combines mtDNA, Y-chromosome and autosomal SNPs in a single, SNaPshot-based assay.

Competing interests

The authors declare no competing interests.

Acknowledgements

We thank Leanne van Weert and Jennifer Young (University of Adelaide) for their technical support during optimisation of the panel; Adrian Linacre (Flinders University) for providing access to equipment; Maria de la Puente (University of Santiago de Compostela) for her assistance in using R for generating figures; and members of the ACAD Thesis Writing Group for critically reviewing previous versions of the manuscript. The research was supported by an Australian Research Council (ARC) Future Fellowship (FT10010008), ARC Discovery Project (DP150101664) and ARC LIEF Project (LE160100154) to JJA.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.fsigen.2018.02.006.

References

[1] C. Philips, Forensic genetic analysis of bio-geographical ancestry, Forensic Sci. Int. Genet. 18 (2015) 49–65.
[2] M. van Oven, M. Vermulen, M. Kayser, Multiplex genotyping system for efficient inference of matrilineal genetic ancestry with continental resolution, Invest. Genet. 2 (2011) 6.
[3] W. Haak, O. Balanovsky, J.J. Sanchez, S. Koshel, V. Zaporozhchenko, C.J. Adler, C.S.I. Sarkissian, G. Brandt, C. Schwarz, N. Nicklisch, V. Dresley, B. Fritsch, E. Balanovska, R. Villems, H. Meller, K.W. Alt, A. Cooper, The genographic consortium, ancient DNA from european early neolithic farmers reveals their near eastern affinities, PLoS Biol. 8 (2010) e1000536.
[4] M. van Oven, A. Ralf, M. Kayser, An efficient multiplex genotyping system approach for detecting the major worldwide human Y-chromosome haplogroups, Int. J. Legal Med. 125 (2011) 879–885.
[5] M. de la Puente, C. Santos, M. Fondevila, L. Manzo, The EUROFORGEN-NoE Consortium, A. Carracedo, M.V. Lareu, C. Philips, The global AIMs nano set: a 31-plex SNaPshot assay of ancestry-informative SNPs, Forensic Sci. Int: Genet. 22 (2016) 81–88.
[6] M. Fondevila, C. Philips, C. Santos, A. Freire Aradas, P.M. Vallone, J.M. Butler, M.V. Lareu, A. Carracedo, Revision of the SNPforID 34-plex forensic ancestry test: assay enhancements standard reference sample genotypes and extended population studies, Forensic Sci. Int.: Genet. 7 (2013) 63–74.
[7] J.J. Sanchez, C. Phillips, C. Børsting, K. Balogh, M. Bogus, M. Fondevila, C.D. Harrison, E. Musgrave-Brown, A. Salas, D. Syndercombe-Court, P.M. Schneider, A. Carracedo, N. Morling, A multiplex assay with 52 single nucleotide polymorphisms for human identification, Electrophoresis 27 (2006) 1713–1724.
[8] S. Walsh, L. Chaitanya, L. Clarisse, L. Wirken, J. Draus-Barini, L. Kovatsi, H. Maeda, T. Ishikawa, T. Sijen, P. de Knijff, W. Branicki, F. Liu, M. Kayser, Developmental validation of the HIrisPlex system: DNA-based eye and hair colour prediction for forensic and anthropological usage, Forensic Sci. Int.: Genet. 9 (2014) 150–161.
[9] ForenSeq™ DNA Illumina Signature Prep Guide, (2014) August.
[10] S.B. Seo, J.L. King, D.H. Warshauer, C.P. Davis, J. Ge, B. Budowle, Single nucleotide polymorphism typing with massively parallel sequencing for human identification, Int. J. Legal Med. 127 (2013) 1079–1086.
[11] J.D. Churchill, S.E. Schmedes, J.L. King, B. Budowle, Evaluation of the illumina beta version ForenSeq DNA signature prep kit for use in genetic profiling, Forensic Sci. Int. Genet. 20 (2016) 20–29.
[12] A.L. Silva, N. Shugarts, J. Smith, A preliminary assessment of the Forenseq FGx system: next generation sequencing of an STR and SNP multiplex, Int. J. Legal Med. 131 (2017) 73.
[13] C. Hussing, C. Borsting, H.S. Mogensen, N. Morling, Testing of the illumina ForenSeq kit, Forensic Sci. Int.: Genet. Supps. Series 5 (2015) e449–e450.
[14] J.J. Sanchez, P. Endicott, Developing multiplexed SNP assays with special reference to degraded DNA templates, Nat. Protoc. 1 (2006) 1370–1378.
[15] J. Ye, G. Coulouris, I. Zaretskaya, I. Cutcutache, S. Rozen, T.L. Madden, Primer-BLAST. A tool to design target-specific primers for polymerase chain reaction, BMC Bioinform. 13 (2012) 134.
[16] C.E. Holleley, P.G. Geerts, Multiplex Manager 1.0: a crossplatform computer program that plans and optimises multiplex PCR, Bio.Tech. 46 (2009) 511–517.
[17] M. van Oven, A. van Geystelen, M. Kayser, R. Decorte, M.H. Larmuseau, Seeing the wood for the trees: a minimal reference phylogeny for the human Y chromosome, Hum. Mutat. 35 (2014) 187–191.
[18] M. van Oven, M. Kayser, Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation, Hum. Mutat. 30 (2009) E386–E394.
[19] The 1000 Genomes Consortium, A global reference for human genetic variation, Nature 526 (2015) 68–74.
[20] H.M. Cann, C. de Toma, L. Cazes, M.F. Legrand, V. Morel, L. Piouffre, J. Bodmer, W.F. Bodmer, B. Bonne-Tamir, A. Cambon-Thomsen, et al., A human genome diversity cell line panel, Science 296 (2002) 261–262.
[21] C. Santos, C. Phillips, A. Gomez-Tato, J. Alvarez-Dios, A. Carracedo, M.V. Lareu, Inference of ancestry in forensic analysis II: analysis of genetic data, in: W. Goodwin (Ed.), Forensic DNA Typing Protocols, Springer, New York, 2016, pp. 256–285.
[22] D. Higgins, A.B. Rohrlach, J. Kaidonis, G. Townsend, J. Austin, Differential nuclear and mitochondrial DNA preservation in post-mortem teeth with implications for forensic and ancient DNA studies, PLoS One 10 (2015) 5.
[23] S. Walsh, F. Liu, K.N. Ballantyne, M. van Oven, O. Lao, M. Kayser, Irisplex: a sensitive DNA tool for accurate prediction of blue and brown eye colour in the absence

of ancestry information, Forensic Sci. Int.: Genet. 5 (2011) 170–180.
[24] E.H. Kim, H.Y. Lee, I.S. Yang, S.E. Jung, W.I. Yang, K.J. Shin, Massively parallel sequencing. of 17 commonly used forensic autosomal STRs and amelogenin with small amplicons, Forensic Sci. Int. Genet. 22 (2016) 1–7.
[25] J.M. Galanter, J.C. Fernandez-Lopez, C.R. Gignoux, J. Barnholtz-Sloan, C. Fernandez-Rozadilla, M. Via, A. Hidalgo-Miranda, A.V. Contreras, L.U. Figueroa, P. Raska, et al., Development of a panel of genome-wide ancestry informative markers to study admixture throughout the americas, PLoS Genet. 8 (2012) e1002554.
[26] C. Santos, C. Phillips, M. Fondevila, R. Daniel, R.A.H. van Oorschot, E.G. Burchard, M.S. Schanfield, L. Suoto, J. Uacyisrael, M. Via, et al., Pacifiplex: an ancestry-informative SNP panel centred on Australia and the Pacific region, Forensic Sci. Int.: Genet. 20 (2016) 71–80.
[27] M. Brion, J.J. Sanchez, K. Balogh, C. Thacker, A. Blanco-Verea, C. Borsting, B. Stradmann-Bellinghausen, M. Bogus, D. Syndercombe-Court, P.M. Schneider, A. Carracedo, N. Morling, Introduction of a single nucleotide polymorphism-based major Y-chromosome haplogroup typing kit suitable for predicting the geographical origins of male lineages, Electrophoresis 26 (2005) 4411–4420.
[28] P. Sulem, D.F. Gudbjartsson, S.N. Stacey, A. Helgason, T. Rafnar, M. Jacobsdottir, S. Steinberg, S.A. Gudjonsson, A. Palsson, G. Thorleifsson, et al., Two newly identified genetic determinants of pigmentation in Europeans, Nat. Genet. 40 (2008) 835–837.
[29] A. Freire-Aradas, Y. Ruiz, C. Phillips, O. Maronas, J. Sochtig, A. Gomez-Tato, J. Alvarez Dios, M. Casares de Cal, V.N. Silbiger, A.D. Luchessi, et al., Exploring iris colour prediction and ancestry in admixed populations of South America, Forensic Sci. Int.: Genet. 13 (2014) 3–9.
[30] A. Wollstein, S. Walsh, F. Liu, U. Chakravarthy, M. Rahu, J.H. Seland, G. Soubrane, L. Tomazzoli, F. Topouzis, J.R. Vingerling, J. Vioque, et al., Novel quantitative pigmentation phenotyping enhances genetic association, epistasis, and prediction of human eye colour, Sci. Rep. 7 (2016) 43359.

Supplementary Data

Supplementary File S1. Details of the 402 reference samples from five population groups across the 1000 Genomes and HGDP-CEPH datasets used for comparison. Genotypes are provided in the strand direction of primers used in the Miniplex for direct comparison (provided as an electronic copy on USB Drive).

Supplementary Table S1. PCR primer information for the Miniplex panel. The code indicates markers in mitochondrial DNA (M), Y chromosome (Y) or autosomal DNA (A) for electropherograms. mtDNA = mitochondrial DNA, Y-chr = Y chromosome DNA, nDNA = nuclear DNA

µM	Code	Marker	Target	Forward Primer	Reverse Primer	Amplicon (bp)	Group	Reference
0.05	M1	3594	mtDNA	AACACAGCAAGACGAGAAGACC	GGACCTGTGGGTTTGTTAGGT	82	L3	Haak et al. 2010
0.05	M2	15301	mtDNA	CCACCCTCACACGATTCTTT	GGTGATTCCTAGGGGGTTGT	119	N	van Oven 2011a
0.05	M3	10400	mtDNA	GCCCTAAGTCTGGCCTATGA	TGAGTCGAAATCATTCGTTTT	90	M	van Oven 2011a
0.15	M4	12705	mtDNA	CCCAAACATTAATCAGTTCTTCAA	TCTCAGCCGATGAACAGTTG	102	R	van Oven 2011a
0.04	M5	5178	mtDNA	ACCCTACTACTATCTCGCACCTGA	CTAGGGAGAGGAGGGTGGAT	76	D	Haak et al. 2010
0.1	Y1	M174	Y-chr	ATGTATCAAATCGCTTCTCTGAATAC	CAAATGCACCCCTCACTTCT	66	D	Haak et al. 2010
0.1	Y2	M96	Y-chr	TGAGCTGTGATGTGTAACTTGG	CACCCACTTTGTTGCTTTGT	117	E	van Oven 2011b
0.15	Y3	M216	Y-chr	CCTCAACCAGTTTTTATGAAGCTA	TTCTAAATCTGAATTCTGACACTGC	102	C	van Oven 2011b
0.1	Y4	M412	Y-chr	GGCACTCCTCCGTCATCTT	GGTGAAGTGGACCCTATCCA	114	R	van Oven 2011b
0.15	Y5	M175	Y-chr	CCCAAATCAACTCAACTCCAG	TTCTACTGATACCTTTGTTTCTGTTCA	101/96	O	van Oven 2011b
0.1	A1	rs9908046	nDNA	CCTTGGCATGTTCCTCTCTC	TCAGAGGAATTAGAAAGGCCTAAA	105	Oceania	de la Puente et al. 2016
0.1	A2	rs1557553	nDNA	TAATACAAGAGCCGCCTGGA	CTTGCAAGGAACTGCAGCTAT	79	Americas	de la Puente et al. 2016
0.1	A3	rs2814778	nDNA	AACCTGATGGCCCTCATTAG	ATGGCACCGTTTGGTTCAG	102	Africa	Fondevila et al. 2013
0.15	A4	rs1426654	nDNA	AATTCAGGAGCTGAACTGCC	TGTTCAGCCCTTGGATTGTC	74	Europe	Fondevila et al. 2013
0.15	A5	rs3827760	nDNA	GCTCAGCTCCACGTACAAC	CTGTCATGCCCCCAATCTC	101	East Asia	Fondevila et al. 2013

0.15	A6	rs12913832	nDNA	GCCCCTGATGATGATAGCGT	GTGTCTGATCCAAGAGGCGA	77	Blue eyes	Own design
0.15	A7	rs1800407	nDNA	TGAAAGGCTGCCTCTGTTCT	CGATGAGACAGAGCATGATGA	127	Intermediate eyes	Walsh et al. 2011
0.1	A8	rs16891982	nDNA	TCCAAGTTGTGCTAGACCAGA	CGAAAGAGGAGTCGAGGTTG	128	Brown eyes	Walsh et al. 2011

Supplementary Table S2. SBE primer information for the Miniplex panel. **b=** Y bases denote equimolar primers used with C/T at two positions (rs144939741 and rs121908454)

µM	Code	Marker	SBE primer	Size in bp	Nucleotides detected	Detection orientation
0.1	M1	3594	ctctctctctctctctctGAGGCCTAGGTTGAGGTT	37	A/G	Reverse
0.3	M2	10400	ctctctctctctctctctctctctcCGTTTTGTTTAAACTATATACCAATTC	53	G/A	Reverse
0.2	M3	15301	ctctctctctctctctctATTCTTTACCTTTCACTTCATCTT	43	A/G	Forward
0.2	M4	12705	ctctctctctctctctctctctctctTTAATCAGTTCTTCAAATATCTACTCAT	55	T/C	Forward
0.07	M5	5178	tctctctctctctctctctctctctctctctctctctctGGAATTAAGGGTGTTAGTCATGTTA	65	G/T	Reverse
0.08	Y1	M174	ctctctctctctctctctctctcATACCTTCTGGAGTGCCC	42	T/C	Forward
0.3	Y2	M96	ctctctctctctctctctctctctctctctctctctctctctctctctTGGAAAACAGGTCTCTCATAATA	72	G/C	Forward
0.6	Y3	M216	ctctctctctctctctctctctctctctctctctctctctctctctctcCTGCTAGTTATGTATACCTGTTGAAT	76	G/A	Reverse
0.7	Y4	M412	ctctctctctctctctctctctctctctctctctctctctctctctctctctctctctctctctctGGGTACAATCTGATGAGGC	86	G/A	Forward
0.6	Y5	M175	ctctctctctctCACATGCCTTCTCACTTCTC	33	T/A(5del)	Forward
0.4	A1	rs9908046	ctctctctctctctctctctctctctctcGAGGTAGGAGCACCCA	46	G/A	Reverse
0.9	A2	rs1557553	ctctctctctctctctctctctctctctcACCCAAAGCCCCTGGAAAAA	50	C/T	Forward
0.8	A3	rs2814778	ctctctctctctctctctctctctctctctctctctctctctctctctctctctctctctctctctctCTCATTAGTCCTTGGCTCTTA	90	T/C	Forward
1.2	A4	rs1426654	ctctctctctctctctctctctctctctctctctctctctctctctctctCGCTGCCATGAAAGTTG	68	C/T	Reverse
0.8	A5	rs3827760	ctctctctctctctctctctctctctctctctctctctctctctctctctctctctctctctctctctctctGGYGCCAYGTTTTCACA**b**	90	A/G	Forward
1	A6	rs12913832	ctctctctctctctctctctctctctctctctctctctctctctctctctctctctctctctctTAGCGTGCAGAACTTGACA	84	T/C	Forward
1.1	A7	rs1800407	ctctctctcCCCACACCCGTCCC	24	C/T	Reverse
0.85	A8	rs16891982	ctctctctctcAAACACGGAGTTGATGCA	30	G/T	Forward

Supplementary Table S3. DNA concentration of high-quality reference samples as measured by Qubit High Sensitivity assay

Sample	DNA Concentration (ng/uL)
ACAD5	1.13
ACAD17	0.448
ACAD34	0.378
ACAD22	0.743
ACAD35	0.692
ACAD8	1.48
ACAD1	0.130
ACAD32	0.941
ACAD29	0.420
ACAD33	1.44

Supplementary Table S4. Sample information for low quality degraded DNA human teeth samples. DNA concentration values were taken from Qubit High Sensitivity assay

Sample	Sex	Decomposition Time (months)	Nuclear DNA concentration (ng/uL)	STR success (%)
401	Male	0	36.2	100
411	Female	0	26.6	100
412	Female	0	2.31	100
491	Female	1	0.586	100
4183	Male	1	0.236	60
4104	Male	1	0.638	100
4113	Female	1	0.177	25
4114	Female	2	0.0560	0
4186	Female	2	0.477	100
4203	Male	2	0.102	50
4181	Male	2	4.73	90
441	Male	4	0.191	45
4178	Female	4	0.167	0
4120	Male	4	1.03	75
435	Female	4	0.291	70
409	Female	8	0.723	60
498	Male	8	0.552	85
402	Male	8	0.684	15
489	Female	16	0.0620	0
496	Female	16	0.065	15
425	Male	16	0.055	30

Supplementary Table S5. Two intermediate likelihood ratios of Snipper inferred ancestry for self-declared high-quality reference samples

Sample	Self-declared ancestry	Snipper inferred ancestry	Intermediate likelihood ratios from Snipper
ACAD5	East Asian	East Asian	15.71 times more likely EAS than AMR and 84.59 times more likely EAS than EUR
ACAD17	Native American	Native American	162 times more likely AMR than EUR and 6647.93 times more likely AMR than OCE
ACAD34	European	European	36,014 times more likely EUR than EAS and 1,980,084 times more likely EUR than OCE
ACAD22	African	African	22,676,802 times more likely AFR than EAS and 135,360,615 times more likely AFR than AMR
ACAD35	East Asian	East Asian	1,484 times more likely EAS than OCE and 1,266,964,856 times more likely EAS than EUR
ACAD8	European	European	4,404,547 times more likely EUR than AMR and 202,074,702 times more likely EUR than AFR
ACAD1	European	European	4,404,547 times more likely EUR than AMR and 202,074,702 times more likely EUR than AFR
ACAD32	European	European	4,404,547 times more likely EUR than AMR and 202,074,702 times more likely EUR than AFR
ACAD33	East Asian	East Asian	618,777 times more likely EAS than OCE and 730,982,443 times more likely EAS than EUR
ACAD29	European	European	4,404,547 times more likely EUR than AMR and 202,074,702 times more likely EUR than AFR

Supplementary Table S6. Two intermediate likelihood ratios of Snipper inferred ancestry for degraded teeth samples

Sample	Snipper inferred ancestry	Intermediate likelihood ratios from Snipper
401	European	4,404,547 times more likely EUR than AMR and 13,071,541 times more likely EUR than OCE
411	European	297,099 times more likely EUR than EAS and 42,559,770 times more likely EUR than AMR
412	European	4,404,547 times more likely EUR than AMR and 13,071,541 times more likely EUR than OCE
491	European	4,404,547 times more likely EUR than AMR and 13,071,541 times more likely EUR than OCE
4183	European	4,404,547 times more likely EUR than AMR and 13,071,541 times more likely EUR than OCE
4104	European	4,404,547 times more likely EUR than AMR and 13,071,541 times more likely EUR than OCE
4113	European	53,991 times more likely EUR than AMR and 289,981 times more likely EUR than EAS
4114	European	1,291 times more likely EUR than EAS and 8,710,694 times more likely EUR than OCE
4186	European	4,404,547 times more likely EUR than AMR and 13,071,541 times more likely EUR than OCE
4203	European	320,812 times more likely EUR than EAS and 9,873,955 times more likely EUR than OCE
4181	European	4,404,547 times more likely EUR than AMR and 13,071,541 times more likely EUR than OCE
441	European	320,812 times more likely EUR than EAS and 9,873,955 times more likely EUR than OCE
4178	European	1,167 times more likely EUR than EAS and 3,299 times more likely EUR than OCE
4120	American	720 times more likely AMR than OCE and 44,677 times more likely AMR than EUR
435	European	1,291 times more likely EUR than EAS and 8,934,100 times more likely EUR than OCE
409	European	320,812 times more likely EUR than EAS and 9,873,955 times more likely EUR than OCE
498	European	4,404,547 times more likely EUR than AMR and 13,071,541 times more likely EUR than OCE
402	NA	Could not be classified
488	European	4,404,547 times more likely EUR than AMR and 13,071,541 times more likely EUR than OCE
415	European	4,404,547 times more likely EUR than AMR and 13,071,541 times more likely EUR than OCE
489	European	1,291 times more likely EUR than EAS and 8,710,694 times more likely EUR than OCE
496	European	5,128 times more likely EUR than AFR and 290,090 times more likely EUR than EAS
425	NA	Could not be classified

Chapter 3

A custom hybridisation enrichment forensic intelligence panel to infer biogeographic ancestry, hair and eye colour and Y-chromosome lineage

Manuscript prepared for submission

Statement of Authorship

Title of Paper	A custom hybridisation enrichment forensic intelligence panel to infer biogeographic ancestry, hair and eye colour and Y-chromosome lineage
Publication Status	☐ Published ☐ Accepted for Publication ☐ Submitted for Publication ☑ Unpublished and Unsubmitted w ork w ritten in manuscript style
Publication Details	Manuscript prepared for publication

Principal Author

Name of Principal Author (Candidate)	Felicia Bardan
Contribution to the Paper	Conceived the study and helped collect the samples. Collected, analysed and interpreted the data, drafted the manuscript and produced the figures.
Overall percentage (%)	70%
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.
Signature	Date 13/11/18

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

i. the candidate's stated contribution to the publication is accurate (as detailed above);

ii. permission is granted for the candidate in include the publication in the thesis; and

iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author	Denice Higgins
Contribution to the Paper	Designed the baits and helped conceive the study. Helped collect the samples, helped analyse the data, revised the manuscript
Signature	Date 09/11/2018

Name of Co-Author	Jeremy J Austin
Contribution to the Paper	Helped conceive the study and assisted in revising manuscript
Signature	Date 14/11/18

A custom hybridisation enrichment forensic intelligence panel to infer biogeographic ancestry, hair and eye colour, and Y chromosome lineage.

Felicia Bardan[1], Denice Higgins[1], Jeremy J. Austin[1]

[1]Australian Centre for Ancient DNA, School of Biological Sciences, University of Adelaide, Adelaide, South Australia 5005, Australia

*corresponding author: felicia.bardan@adelaide.edu.au

Abstract

Massively parallel sequencing provides new opportunities to obtain genetic data for hundreds of loci in a single assay for various types of forensic testing. However, available commercial kits are based on an initial PCR amplification of short-to-medium sized targets which creates problems for highly degraded DNA. Development and optimisation of large PCR multiplexes also limits the ability to create custom panels that target different suites of markers for identity, biogeographic ancestry, phenotype and lineage markers (Y-chromosome and mtDNA). Hybridisation enrichment, an alternative approach to target enrichment prior to sequencing, uses biotinylated probes to bind to target DNA and has proven successful on degraded and ancient DNA. We developed a customisable hybridisation capture method, whereby the use of individually mixed baits allows the user the possibility to tailor target enrichment to specific questions of interest. To allow collection of forensic intelligence data, we assembled and tested a custom panel of hybridisation baits to examine biogeographic ancestry, phenotype and paternal lineage. We demonstrate the ability of this panel to infer biogeographic ancestry, hair and eye colour, and paternal lineage (and sex) on modern male and female samples with a range of self-declared ancestries and hair/eye colour combinations. The panel correctly estimated biogeographic ancestry in 9/12 samples (75%) but detected European admixture in three individuals from regions with admixed demographic history. Hair and eye colour were predicted correctly in 83% and 92% of samples respectively, where intermediate eye colour and blond hair were problematic to predict. Analysis of Y-chromosome SNPs correctly assigned sex and paternal haplogroups, the latter complementing and supporting biogeographic ancestry predictions. Overall, we demonstrate the utility of this hybridisation enrichment approach to forensic intelligence testing using a combined suite of biogeographic ancestry, phenotype and Y-chromosome SNPs for comprehensive biological profiling.

Keywords: forensic intelligence; SNPs; biogeographic ancestry; phenotype; hybridisation enrichment

Introduction

The genotyping of single nucleotide polymorphisms (SNPs) is a useful alternative to complement conventional Short Tandem Repeat (STR) typing for problematic forensic samples. SNP genotyping has the advantages of targeting shorter DNA fragments than STR typing, and a broader range of biological information can be gathered (Budowle & van Daal 2008). For example, SNPs demonstrating low population heterogeneity and a high degree of polymorphism are useful for informing individual identity in the same manner as STR typing (Musgrave-Brown *et al.* 2007). Conversely, ancestry informative SNPs have low heterozygosity and highly contrasting allele frequencies between populations, making them useful for predicting an individual's biogeographic ancestry (Phillips 2015). SNP variation in pigmentation genes can also be useful for inferring external visible characteristics such as hair and eye colour (Walsh et al. 2014).

Hybridisation enrichment combined with massively parallel sequencing (MPS) has been explored recently for forensic identification purposes targeting mitochondrial DNA (mtDNA) and nuclear SNPs (Templeton *et al.* 2013; Bose *et al.* 2018). These studies demonstrate the utility of this method for generating whole mitochondrial genomes and hundreds of SNPs from forensic samples. Commercial MPS panels using standard PCR-based target enrichment have also been developed to genotype many forensically relevant markers (de la Puente et al. 2017; Meiklejohn et al. 2017; Xavier et al. 2017). However, these methods present a largely inflexible approach to SNP typing where targets are amplified and captured in pre-mixed multiplexes.

Here we test a customisable hybridisation enrichment SNP panel for interrogation of biogeographic ancestry, phenotype and Y-chromosome (Y-chr) SNPs. This panel has the possibility of customisation of SNPs targeted by the use of individually mixed baits which can be prepared in any number and combination with minimal optimisation (some hybridisation enrichment panels target up to ten thousand SNPs; Soubrier *et al.* 2016) compared to multiplex PCP strategies which require extensive primer design and laboratory optimisation, making target customisation a complex task. The panel distinguishes between five continental population groups: Africa, Native America, East Asia, Europe and Oceania, allows sex determination and the prediction of hair and eye colour. For use in Australian forensic applications, this panel includes markers for differentiating Oceanian ancestry from other populations with the inclusion of biogeographic and Y-chr SNPs informative for Oceanian/Australian Aboriginal ancestry. The selection of the Y-chr SNPs provides a global

coverage of major Y-chr haplogroups. It also provides further resolution of some Y-chr haplogroups which have distributions across multiple geographic regions, such as haplogroup C, E, O and R which can be found across Africa, Europe and Asia (Cruciani *et al.* 2007; van Oven *et al.* 2013; Underhill *et al.* 2014). The customisable hybridisation enrichment panel targets 124 SNP markers including: 67 biogeographic ancestry informative markers (including four tri-allelic markers), 23 phenotype markers from the HIrisPlex panel (with one SNP overlapping as a biogeographic ancestry SNP)(Walsh *et al.* 2014), and 35 Y-chr SNPs. This study forms an initial evaluation of the custom panel and assesses prediction accuracy on a set of modern human DNA samples with known biogeographic ancestry, phenotype and sex. This forensic assay has the potential to provide investigative leads for cases involving missing persons, historical human remains, and trace samples where DNA typing can be augmented with information regarding the physical appearance and biogeographic ancestry of the person from whom the sample originated.

Materials and Methods

SNP selection

Initially we selected 53 autosomal SNPs included in two existing forensic panels - the 31 biogeographic ancestry SNPs in the Global AIMs Nano (GNano) set (de la Puente *et al.* 2016) and 23 of the 24 SNPs in the HIrisPlex (Walsh *et al.* 2014) panel for hair and eye colour prediction. The HIrisPlex N29insA indel was excluded and SNP rs16891982 is shared between the Global AIMs Nano and HIrisPlex panels. Subsequently we selected 22 biogeographic ancestry SNPs, including two X-chromosome SNPs, from the PacifiPlex ancestry panel (Santos *et al.* 2016) (seven SNPs overlap with the GNano set), and 15 biogeographic ancestry SNPs based on their appearance in other forensic ancestry panels, and with the highest remaining divergence values across continental groups from Phillips *et al.* (2014), Kosoy *et al.* (2008), Daya *et al.* (2013), and Daca-Roszak *et al.* (2016). Thirty-five Y-chromosome SNPs were selected from Karafet *et al.* (2008, 2010, 2015), Lao *et al.* (2010), van Oven *et al.* (2011, 2012, 2013b), Valverde *et al.* (2013), Park *et al.* (2013), and Hudjashov *et al.* (2007). The Y-chromosome SNPs were chosen to infer major worldwide macrohaplogroups, and additional SNP for dissecting Y-haplogroups with broad distributions (such as C, E, O and R) into sub-haplogroups with more restricted geographical affiliations (Figure 2). For application to questions of patrilineal lineage in Australia, further resolution of haplogroup C distinguishes between subgroups common across East Asia, South East Asia

and Oceania versus those specific to Australian Aboriginals (e.g. C-M347)(Hudjashov *et al.* 2007; Zhong *et al.* 2010; van Oven *et al.* 2011; Taylor *et al.* 2012; Naitoh *et al.* 2013; Bergström *et al.* 2016). In total we selected 125 SNPs (67 ancestry; Table 1, 23 phenotype; Table 2, 35 Y-chr; Table 3, with one SNP shared between ancestry and phenotype). MtDNA SNPS were not targeted in this panel given our laboratory has previously developed a whole mtDNA hybridisation enrichment strategy in Templeton *et al.* (2013).

Table 1. Details of the 67 biogeographic ancestry SNPs included in the hybridisation enrichment panel. Note that EUR informative SNP at rs16891982 is also included in the phenotype SNPs. Ancestry groups are: AFR-African, AMR-Native American, EAS=East Asian, EUR-European, OCE-Oceanian. Tri-allelic SNPs are ancestry informative but also serve to monitor for contamination from >1 DNA donor.

rs Number	Chr.	Position (GRCh37/hg19)	Ancestry Group	Reference
rs2139931	1	84590527	OCE	de la Puente *et al.* 2016/Santos *et al.* 2016
rs2814778	1	159174683	AFR	de la Puente *et al.* 2016
rs4657449	1	165465281	EAS	de la Puente *et al.* 2016
rs12142199	1	1249187	EUR	de la Puente *et al.* 2016
rs12402499	1	101528954	AMR	de la Puente *et al.* 2016
rs647325	1	18170886	AMR	Kosoy *et al.* 2008
rs2184030	1	206667441	Tri-allelic	Santos *et al.* 2016
rs16830500	2	152814129	OCE	Phillips *et al.* 2014
rs3827760	2	109513601	EAS	de la Puente *et al.* 2016/Santos *et al.* 2016
rs10183022	2	237481969	OCE	Santos *et al.* 2016
rs820371	3	123404711	EUR	Phillips *et al.* 2014
rs6437783	3	108172817	EAS	de la Puente *et al.* 2016
rs9809818	3	71480566	OCE	de la Puente *et al.* 2016/Santos *et al.* 2016
rs12498138	3	121459589	AMR	de la Puente *et al.* 2016
rs4683510	3	140285115	EAS	Santos *et al.* 2016
rs7623065	3	22385375	OCE	Santos *et al.* 2016
rs10012227	4	18637315	AMR	Phillips *et al.* 2014
rs1229984	4	100239319	EAS	de la Puente *et al.* 2016
rs4540055	4	38803255	Tri-allelic	de la Puente *et al.* 2016
rs1509524	4	125455038	OCE	Santos *et al.* 2016
rs6875659	5	175158653	AFR	Phillips *et al.* 2014
rs16891982	5	33951693	EUR	de la Puente *et al.* 2016/Walsh *et al.* 2014
rs4704322	5	75822474	EAS	Santos *et al.* 2016
rs6886019	5	170245846	OCE	Santos *et al.* 2016
rs10455681	6	69802502	OCE	Santos *et al.* 2016
rs2080161	7	13331150	AMR	de la Puente *et al.* 2016
rs798949	7	120765954	OCE	Santos *et al.* 2016
rs1871534	8	145639681	AFR	de la Puente *et al.* 2016
rs2409722	8	11039816	OCE	Santos *et al.* 2016
rs7832008	8	98358246	OCE	Santos *et al.* 2016
rs2789823	9	136769888	AFR	de la Puente *et al.* 2016
rs10811102	9	1911291	OCE	Santos *et al.* 2016
rs10970986	9	32453278	OCE	Santos *et al.* 2016
rs16913918	9	3074359	EUR	Daya *et al.* 2013
rs7084970	10	119750413	EUR	Phillips *et al.* 2014
rs4749305	10	28391596	EUR	de la Puente *et al.* 2016
rs2274636	10	27443012	OCE	Santos *et al.* 2016
rs174570	11	61597212	AMR	Phillips *et al.* 2014
rs3751050	11	9091244	OCE	de la Puente *et al.* 2016/Santos *et al.* 2016

rs5030240	11	32424389	Tri-allelic	de la Puente *et al.* 2016
rs1924381	13	72321856	EUR	Phillips *et al.* 2014
rs9522149	13	111827167	EUR	de la Puente *et al.* 2016
rs721367	13	95546650	EAS	Santos *et al.* 2016
rs730570	14	101142890	EUR	Phillips *et al.* 2014
rs7151991	14	32635572	AMR	Phillips *et al.* 2014
rs10483251	14	21671277	AMR	de la Puente *et al.* 2016
rs12434466	14	97324289	EAS	Santos *et al.* 2016
rs1834640	15	48392165	EUR	Daca-Roszak *et al.* 2016
rs12594144	15	64161351	EAS	de la Puente *et al.* 2016
rs1426654	15	48426484	EUR	de la Puente *et al.* 2016
rs3784651	15	94925273	OCE	Santos *et al.* 2016
rs6494411	15	63835861	EAS	Santos *et al.* 2016
rs881929	16	31079371	EAS	Phillips *et al.* 2014
rs17822931	16	48258198	EAS	de la Puente *et al.* 2016
rs16946159	16	48459558	OCE	Santos *et al.* 2016
rs4792928	17	42105174	AMR	de la Puente *et al.* 2016
rs8072587	17	19211073	EUR	de la Puente *et al.* 2016
rs9908046	17	53563782	OCE	de la Puente *et al.* 2016/Santos *et al.* 2016
rs1369290	18	67691520	AFR	Phillips *et al.* 2014
rs310644	20	62159504	AFR	Phillips *et al.* 2014
rs2069945	20	33761837	Tri-allelic	de la Puente *et al.* 2016/Santos *et al.* 2016
rs6054465	20	6673018	OCE	de la Puente *et al.* 2016/Santos *et al.* 2016
rs715605	22	30640308	OCE	de la Puente *et al.* 2016
rs1557553	22	44760984	AMR	de la Puente *et al.* 2016
rs8137373	22	41729216	AMR	de la Puente *et al.* 2016
rs4892491	X	73422412	EAS	Santos *et al.* 2016
rs11156577	X	153660041	OCE	Santos *et al.* 2016

Table 2. Details of the 23 phenotype (hair and eye colour) SNPs included in the hybridisation enrichment panel. Note that the SNP at rs16891982 is also included in the ancestry SNPs.

rs Number	Chr.	Position (GRCh37/hg19)	Reference
rs16891982	5	33951693	de la Puente *et al.* 2016/ Walsh *et al.* 2014
rs28777	5	33958959	Walsh *et al.* 2014
rs4959270	6	457748	Walsh *et al.* 2014
rs12203592	6	396321	Walsh *et al.* 2014
rs683	9	12709305	Walsh *et al.* 2014
rs1042602	11	88911696	Walsh *et al.* 2014
rs1393350	11	89011046	Walsh *et al.* 2014
rs12821256	12	89328335	Walsh *et al.* 2014
rs2402130	14	92801203	Walsh *et al.* 2014
rs12896399	14	92773663	Walsh *et al.* 2014
rs12913832	15	28365618	Walsh *et al.* 2014
rs1800407	15	28230318	Walsh *et al.* 2014
rs1805005	16	89985844	Walsh *et al.* 2014
rs1805006	16	89985918	Walsh *et al.* 2014
rs2228479	16	89985940	Walsh *et al.* 2014
rs11547464	16	89986091	Walsh *et al.* 2014
rs1805007	16	89986117	Walsh *et al.* 2014
rs201326893	16	89986122	Walsh *et al.* 2014
rs1110400	16	89986130	Walsh *et al.* 2014
rs1805008	16	89986144	Walsh *et al.* 2014
rs885479	16	89986154	Walsh *et al.* 2014
rs1805009	16	89986546	Walsh *et al.* 2014
rs2378249	20	33218090	Walsh *et al.* 2014

Table 3. Details of the 35 Y-chromosome SNPs included in the hybridisation enrichment panel. P203 (also known as M307) is informative for both O and I haplogroups as described in ISOGG 2018 v 13.256 and Karafet *et al.* (2008). Haplogroup is defined by diagnostic SNP.

rs Number (mutation name)	Position (GRCh37/hg19)	Y-chr haplogroup	Reference
rs2032595 (M168)	14813991	CDEF-M168	Valverde *et al.* 2013
rs3848982 (M145)	21717208	DE-M145	Valverde *et al.* 2013
rs2032602 (M174)	14954280	D-M174	Lao *et al.* 2010
rs371443469 (V36)	6814246	E-V36	van Oven *et al.* 2013
rs9306841 (M96)	21778998	E-M96	van Oven *et al.* 2011
rs9786025 (P170)	15021522	E-P170	Valverde *et al.* 2013
rs2032666 (M216)	15437564	C-M216	van Oven *et al.* 2011
rs35284970 (M130)	2734854	C-M130	Valverde *et al.* 2013
rs2032668 (M217)	15437333	C-M217	Park *et al.* 2013
rs868363758 (M347)	2877479	C-M347	Hudjasov *et al.* 2007
rs9786706 (U13)	14698928	G-U13	van Oven *et al.* 2013
rs2032636 (M201)	15027529	G-M201	Valverde *et al.* 2013
rs13447371 (M282)	21764431	H-M282	van Oven *et al.* 2011
rs2032673 (M69)	21894058	H-M69	Valverde *et al.* 2013
rs17250163 (P126)	21225770	IJ-P126	Valverde *et al.* 2013
rs9341301 (M258)	15023364	I-M258	Valverde *et al.* 2013
rs13447352 (M304)	22749853	J-M304	van Oven *et al.* 2011
rs9341313 (M267)	22741818	J-M267	van Oven *et al.* 2013
rs3900 (M9)	21730257	KLT-M9	Valverde *et al.* 2013
rs3902 (M11)	21730647	L-M11	Valverde *et al.* 2013
rs9341308 (M272)	22738775	T-M272	Valverde *et al.* 2013
rs2033003 (M526)	23550924	K-M526	van Oven *et al.* 2011
n/a (P308)	15409573	S-P308	Karafet *et al.* 2015
n/a (P256)	8685230	M-P256	Karafet *et al.* 2008
rs2032631 (M45)	21867787	P-M45	Valverde *et al.* 2013
rs8179021 (M242)	15018582	Q-M242	Valverde *et al.* 2013
rs2032658 (M207)	15581983	R-M207	Valverde *et al.* 2013
rs17250535 (M420)	23473201	R-M420	van Oven *et al.* 2013
rs9786184 (M343)	2887824	R-M343	van Oven *et al.* 2013
rs9786153 (M269)	22739367	R-M269	van Oven *et al.* 2013
rs9786140 (M412)	8502236	R-M412	van Oven *et al.* 2011
rs9341278 (M231)	15469724	N-M231	Valverde *et al.* 2013
rs13447361 (M324)	2821786	O-M324	van Oven *et al.* 2012
rs11575897 (M176)	2655180	O-M176	van Oven *et al.* 2012
rs13447354 (P203)	22750951	O-P203/I-P203	Karafet *et al.* 2010

Bait design

Single-stranded DNA bait sequence design was performed using the Integrated DNA Technologies (IDT) Target Capture Probe Design Tool (https://sg.idtdna.com/site/order/ngs) with probe length set to 120 bp and tiling density at 1x. Bait sequences were searched against the NCBI database using Blast to ensure baits were homologous only to the desired SNP location in the human genome. Sequences with more than one hit were redesigned by moving the start and finish positions. We could not design a unique 120-mer bait around one SNP (rs12405776 from the PacifiPlex panel) and so this SNP was excluded. As GC content has been shown to influence the performance of baits (Tewhey *et al.* 2009; Aird *et al.* 2011;

Dabney & Meyer 2012), we aimed for a GC content of 40 – 60 %. Minor adjustments in the start and finish positions of the bait sequences were made to aid improve GC content, however the final GC content varied from 26.7 – 68.3 %. Three (rs1805005, rs1805006, rs2228479) and six (rs11547464, rs1805007, rs201326893, rs1110400, rs1805008 and rs885479) phenotype SNPs from the HIrisPlex panel occur in short (97 and 64 bp, respectively) genomic segments, so we designed single baits that included these linked markers. Thus, the final 124 SNP set was covered by a total of 117 baits (Supplementary Table S1, S2, S3). Baits were synthesised as single-tiled 5' biotinylated 120-mer DNA oligonucleotides (xGen Lockdown Probes) by IDT and rehydrated to a concentration of 1mM in 1xTE buffer. The final bait pool was created by combining the individual baits at a final concentration of 100 aM/uL per bait in 1xTE buffer.

Reference population differentiation analysis

To assess the population differentiation power of the 67 ancestry-informative SNP markers in the hybridisation panel, Shannon's Divergence values were calculated from a reference population set comprising genotypes from 368 individuals from African (AFR, n = 99), East Asian (EAS, n = 89), European (EUR, n = 88), Native American (AMR, n = 64), and Oceanian (OCE, n = 28) populations (see Supplementary File S4). Values were produced for each SNP using the Bayesian classifier Snipper portal 'cross-validation' analysis (mathgene.usc.es/snipper/) from comparisons of one population against all others (e.g. AFR vs all non-AFR populations), and for pairwise population analyses. Shannon's Divergence values were converted to the Rosenberg's 'informativeness for assignment' statistic (I_n) by multiplication with 0.693 (Rosenberg *et al.* 2003). Final population-specific divergence (PSD) values for each ancestral population group was obtained from the cumulative divergence (I_n) values. The Snipper portal was also used for cross-validating reference population data for ancestry assignment accuracy. STRUCTURE (with no POPFLAG) was also applied to the reference population dataset in Supplementary File S4 to further assess the population differentiation power of the 67 SNPs in the enrichment panel.

To assess the reference population differentiation of the 67 SNPs in the custom panel in comparison to current SNaPshot SNP-typing technology recently implemented in our laboratory, an evaluation against the Global AIMs Nano set (31 SNPs) was conducted (de la Puente *et al.* 2016). PCA analysis was applied to the same reference population dataset described above.

Test Samples and DNA extractions

High quality DNA was obtained via buccal swabs collected from twelve human donors with informed consent in accordance with ethics approval from the University of Adelaide Human Research and Ethics Committee (H-2016-218). Donors had a range of self-declared biogeographic ancestry (Europe, Africa, Native America, East Asia), sex, and included people with varying combinations of blue, intermediate and brown eyes and brown, black, red and blond hair (Table 4). No samples from donors with Oceanian ancestry were available.

Table 4. DNA samples used in this study with a range of self-declared biogeographic ancestry, sex, and reported hair and eye colour.

Sample	Sex	Self-declared ancestry	Region	Reported Hair Colour	Reported Eye Colour
S5	Male	European	Western Europe	Brown	Brown
S12	Male	East Asian	South-East Asia	Black	Brown
S2	Male	Native American	Central America	Brown	Brown
S3	Male	African	Sub-Saharan Africa	Black	Brown
S6	Male	European	Western Europe	Blond	Blue
S9	Male	African	North Africa	Black	Brown
S1	Female	European	Western Europe	Red	Blue
S4	Female	European	Western Europe	Blond	Blue
S7	Female	European	Western Europe	Red	Intermediate
S8	Female	East Asian	Mainland East Asia	Black	Brown
S10	Female	Native American	Central America	Brown	Brown
S11	Female	European	Western Europe	Blonde	Intermediate

Buccal swabs on FTA card were extracted from a ~5 mm2 punch for each sample (and an extraction blank) using the QIAmp DNA Mini Kit (Qiagen, Hagen, Germany) following the manufacturer's instructions in a final volume of 150 uL. DNA extracts were mechanically fragmented for library preparation using a Covaris S220 Focused-ultrasonicator (Covaris) to ~150 bp following the manufacturer's instructions. Sheared DNA was purified and concentrated to 21 uL using 1.2x volume of Agencourt AMpure XP beads (Beckman-Coulter) according to manufacturer's instructions and then quantified using the Qubit fluorometer double-stranded High Sensitivity assay (Life Technologies).

Library preparation

Twenty microlitres of fragmented genomic DNA from each sample (21 - 95 ng total) was converted into double-stranded Illumina libraries, using truncated Illumina P5 and P7 adapters with dual 7-mer internal barcodes, following the protocol of Meyer et al. (2010)

with minor modifications described by Llamas et al. (2016). Fragment size distribution and DNA concentration were measured on a 2.5% agarose gel against Hyperladder IV (Bioline) and using a Qubit double-stranded DNA High Sensitivity assay (Thermo Fisher).

Hybridisation enrichment and sequencing

Hybridisation enrichment was conducted according to the IDT 'Hybridisation capture of DNA libraries using xGen Lockdown Probes and Reagents, v1' with slight modification as follows (Figure 1). Samples were prepared by combining 197-311 ng of barcoded DNA library with 2.5 µg Human Cot1 DNA (Invitrogen), 2.5 µg Salmon Sperm DNA (Invitrogen), 25 pmol of blocking RNA oligonucleotides (matching the sequence of the truncated Illumina P5 and P7 adapters), and 20 U of SUPERase-In (Ambion). The samples were then dried using a CentriVap Vacuum Concentrator (LABCONCO) at 45°C ~ 30 min. The library DNA was rehydrated in a final volume of 17 µL containing 8.5 uL xGen 2x Hybridisation Buffer, 2.7 uL xGen Hybridisation Buffer Enhancer, 1.8 uL dH20, and 4 µL of custom bait pool then denatured at 95°C for 10 min. A step-down hybridisation reaction over 40 hrs was used with 5 hrs at 65°C, 5 hrs at 60°C, and 30 hrs at 55°C.

For each hybridisation reaction, 30 µL of Dynabeads® MyOne Streptavidin C1 (Life Technologies) was washed twice with 200 uL of xGen 1X Bead Wash Buffer (IDT) at room temperature. The entire hybridisation reaction was transferred to the beads and incubated at 55°C for 45 min (with vortexing every 10 min) to allow binding of the biotin to the streptavidin. The bead-hybridised DNA complex was then washed at 65°C with 100 uL of 1X Wash Buffer 1 (IDT) for 2 min, followed by two washes with 200 uL of 1X Stringent Wash Buffer (IDT) at 65°C for 5 min. A series of RT washes were then performed using 200 µL of 1X Wash Buffer 1 (IDT) for 2 min, once with 200 µL of 1X Wash Buffer 2 (IDT) for 1 min and once with 200 µL of 1X Wash Buffer 3 (IDT) for 30 sec.

After the last wash, beads were resuspended in 16 uL of 10mM Tris (pH8.0) + 0.05% Tween-20. Each enriched library was amplified in five 25 µL reactions containing 1× AmpliTaq Gold buffer, 2.5 mM $MgCl_2$, 250 µM of each dNTP, 0.01 U AmpliTaq Gold (Applied Biosystems), 0.5 µM of primers IS7 and IS8 (Meyer & Kircher 2010) and 3 uL of enriched DNA. Thermocycling was completed at 94°C for 12 min, followed by 14 cycles of 30 s at 94°C, 30 s at 60°C and 45 s at 72°C, followed by a final extension of 10 min at 72°C. Amplification replicates for each sample were pooled and purified using Agencourt AMpure

XP beads (Beckman-Coulter) at 1.2x volume and eluted in 30 uL EB Buffer + 0.05% Tween-20 (Qiagen).

Enriched libraries were then taken through a second round of enrichment using the same protocol above to produce double-enriched DNA libraries. Previous optimisation experiments (data not shown) and the prior development of a mtDNA genome hybridisation enrichment technique within our laboratory has demonstrated improved enrichment of targets when using two rounds of enrichment versus one round (Templeton *et al.* 2013). A final PCR amplification using full-length 7-mer indexed Illumina adapters was performed as described in Meyer & Kircher 2010. Library fragment size distribution and concentration were measured using a Bioanalyzer 2100 (Agilent Technologies) following the manufacturer's instructions. Samples were pooled equimolar for a final concentration of 5 nmol/L prior to sequencing on an Illumina MiSeq at the Australian Genome Research Facility (AGRF) using paired end 150 bp sequencing.

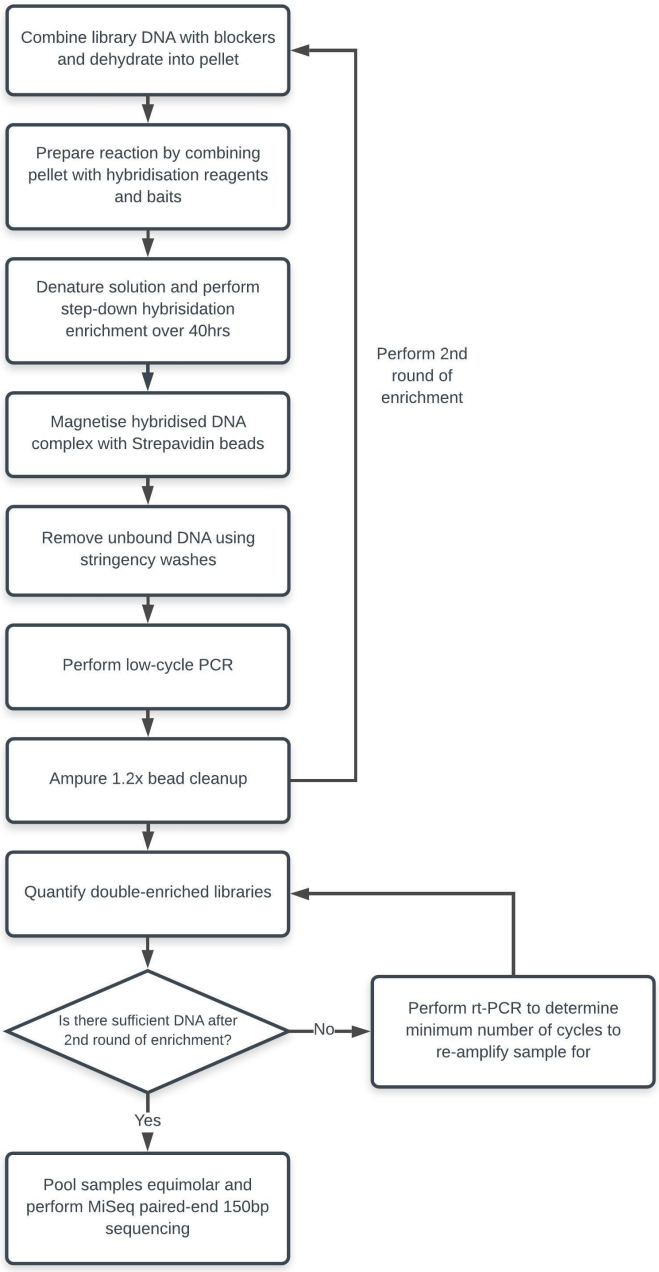


Figure 1. Workflow for the hybridisation enrichment and sequencing process

Data analysis

Following sequencing, reads were filtered according to standard Illumina protocol at AGRF to remove low-quality clusters, de-multiplex by index and Illumina indices were trimmed. Raw Illumina reads were subsequently processed using the PaleoMix v1.0.1 pipeline (Schubert *et al.* 2014). Sequences were de-multiplexed for specific samples using the dual P5/P7 internal barcodes. Adapter removal V2 (Lindgreen 2012) was used to trim adapters, merge paired reads and eliminate all reads shorter than 25 bp long. Collapsed reads were then mapped to the Human Genome GRCh37/hg19 using BWA v0.6.2 (Li & Durbin 2009).

Minimum mapping quality was set to 25 and seeding was disabled. PCR duplicates (mapped reads that start and finish at the same location) were removed to retain only unique reads for genotype calling. SNPs were then called using SAMTools mpileup/bcftools to generate a vcf file. Genotypes for the SNPs of interest were then isolated via interrogation against a custom BED file containing the genomic coordinates of targeted SNP loci.

Biogeographic ancestry prediction

For assignment of biogeographic ancestry, the 67 ancestry informative SNPs from each sample genotype were compared to a reference population set comprising genotypes from 368 individuals from African (AFR, n = 99), East Asian (EAS, n = 89), European (EUR, n = 88), Native American (AMR, n = 64), and Oceanian (OCE, n = 28) populations (Table 5). Reference population genotypes were obtained from the 1000 Genomes Phase II (The 1000 Genomes Project Consortium 2015) and Stanford University HGDP-CEPH (Cann *et al.* 2002) datasets (Supplementary File S4). Ancestry assignment was performed using Snipper (*mathgene.usc.es/snipper/*), with Hardy-Weinberg principle applied. Likelihood ratios (LR) for ancestry classifications were used for direct ancestry estimation, and principle component analysis (PCA) was performed in RStudio (v1.1.442) with the *SNPassoc* package to visually summarise the genetic differences and similarities of the sample genotypes to the reference populations (Gonzalez *et al.* 2007).

Table 5. Populations used as a reference population set for ancestry informative SNPs. AFR: African, AMR: Native American, EAS: East Asian, EUR: Europe, OCE: Oceanian.

Population	N	Data Source	Description
AFR	99	1000 Genomes	ESN: Esan in Nigeria
AMR	64	HGDP-CEPH	22 from Brazil, 35 from Mexico, 7 from Colombia
EAS	89	1000 Genomes	JPT: Japanese in Tokyo
EUR	88	1000 Genomes	GBR: British in England and Scotland
OCE	28	HGDP-CEPH	17 from New Guinea & 11 from Bougainville

Genetic ancestry for each sample was further assessed by applying the admixture model to the ancestry SNP data in STRUCTURE v.2.3.4 (Porras-Hurtado *et al.* 2013). The reference population set described above was used for population membership analysis of the twelve test samples in STRUCTURE. The number of clusters (K) considered in the analysis ranged from 2 - 7. Only results for K=5 are presented as it was identified as the optimal K value from STRUCTURE HARVESTER (Earl & vonHoldt 2012). Analyses with STRUCTURE were performed using the following parameters: five iterations of 100,000 burnin steps and 100,000 MCMC steps, correlated allele frequencies under the Admixture model (including

POPFLAG). POPFLAG uses the population of origin details (specified by the user for reference population samples; POPFLAG = 1) to help infer the ancestry of test samples (POPFLAG = 0) with unknown origin based on allele frequencies. Estimated cluster membership coefficients from STRUCTURE analysis were used to construct population membership bar plots with CLUMPAK v.1.1 (Kopelman *et al.* 2015) for visualisation.

Phenotype prediction

The 23 phenotype SNPs were used to predict hair and eye colour using the prediction model from the HIrisPlex Eye and Hair Colour DNA Phenotyping Webtool (*hirisplex.erasmusmc.nl*) outlined by Walsh *et al.* (2014). SNPs were entered into the interface to generate probabilities of belonging to a particular phenotypic class for hair and eye colour, along with a 'loss in accuracy of prediction' (AUC) value for any missing SNPs using a multinomial logistic regression model.

For eye colour, the current prediction framework specified in Walsh *et al.* (2014) states that the highest p-value indicates most likely eye colour. However, for hair colour, current interpretation guidelines combine the highest p-value approach in conjunction with a step-wise model taking into account the shade probability values (i.e. light or dark) to infer 'most probable hair colour' (see Supplementary Figure 2 in Walsh *et al.* 2014).

Sex, Y Chromosome haplogroup and paternal geographic ancestry

The 35 Y-chr SNP profile generated was used to identify the sex of individuals (females = ancestry and phenotype SNPs but no Y-chr SNPs; males = ancestry, phenotype and Y-chr SNPs), and define the Y haplogroup for each male individual according to diagnostic ancestral and derived SNPs in PhyloTree (ISOGG 2018 v 13.256), Valverde et al. (2013) and Karafet *et al.* (2015). SNPs were checked for phylogenetic sense (i.e. no conflicting haplogroup assignments). Geographical affiliation (Figure 2) was assigned based on the classifications and frequencies in previous studies (Lao *et al.* 2010; van Oven *et al.* 2011; Park *et al.* 2013; Valverde *et al.* 2013; Karafet *et al.* 2015; Nagle *et al.* 2015).

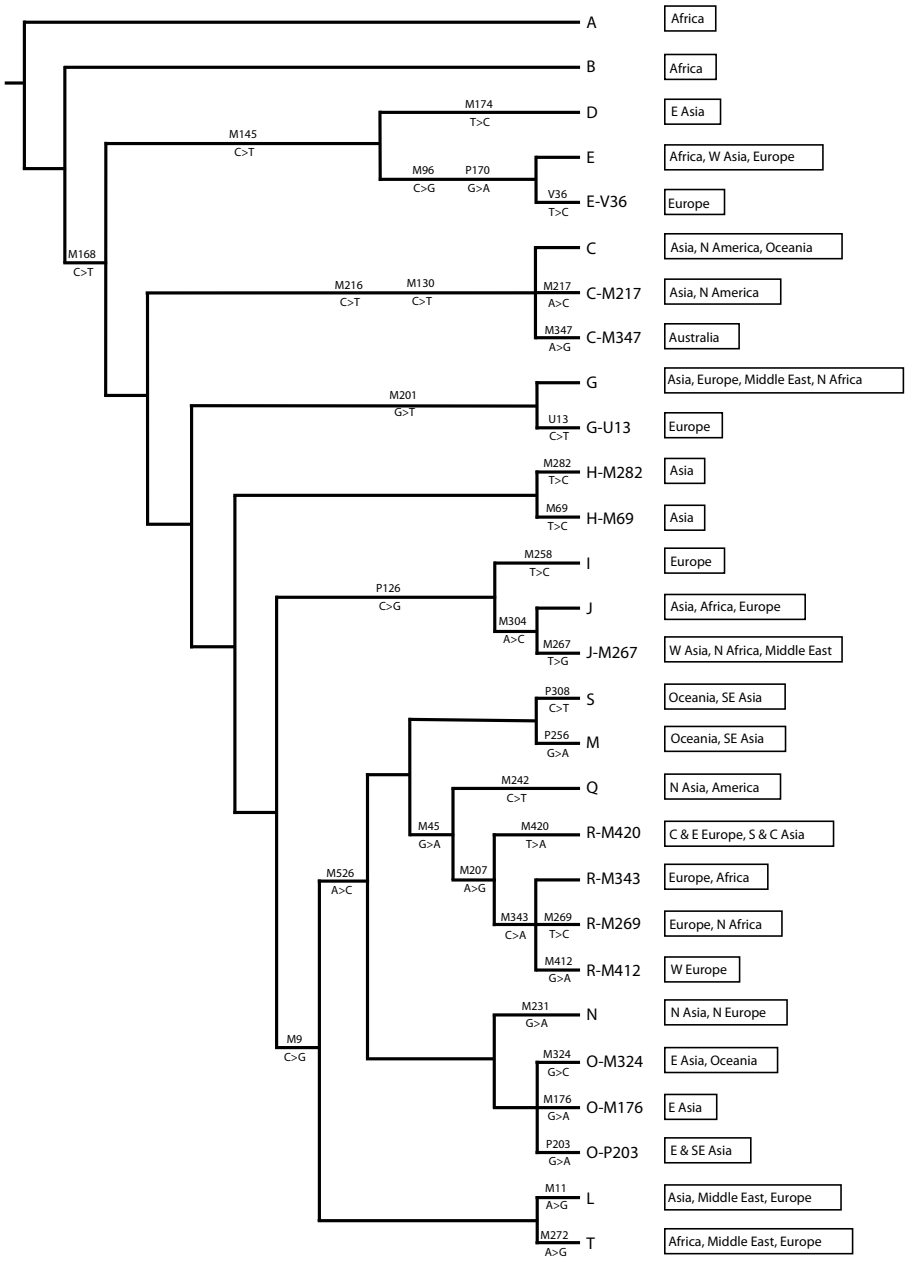


Figure 2. Simplified phylogenetic tree of the 35 Y-chr SNPs included in the custom hybridisation enrichment panel, haplogroups that can be inferred from them, and the main geographical distribution of the haplogroups. Additional SNPs in haplogroups E, C, J, R and O were chosen to distinguish between sub-haplogroups.

Quality control

Extraction blanks were included in each extraction batch. No-template controls during library preparation were included in each experiment (including during library preparation and hybridisation enrichment). All controls were included to monitor potential contamination from exogenous human DNA sources and cross-contamination from other samples. All work prior to library amplification (and before enrichment reactions) was conducted in a geographically separate laboratory to post-amplification laboratories.

Results

Population differentiation of the 67 biogeographic ancestry SNPs in the custom enrichment panel

'One-against-others' comparisons of reference population group data produced cumulative PSD values shown in Table 6. Further details for the divergence values for each SNP are given in Supplementary Table S5.

Table 6. One-against-others population-specific Divergence (PSD) values for each of the five reference population groups used for assessing the population differentiation capabilities of the custom enrichment panel.

Population	Cumulative PSD
AFR	9.77
AMR	7.27
EAS	7.33
EUR	8.05
OCE	9.7

This analysis indicated that the divergence values of EAS, EUR and AMR population groups were lower than the average of 8.43. The highest PSD value belonged to the AFR group. Cross-validation of reference population data showed 100% success for ancestry assignment when using this SNP set across all population groups. Pairwise comparisons indicated populations with the lowest differentiation were EAS from AMR (6.70) and OCE (11.13) populations (Figure 3), supported by the lower divergence values for EAS-informative SNPs (Supplementary Table S5).

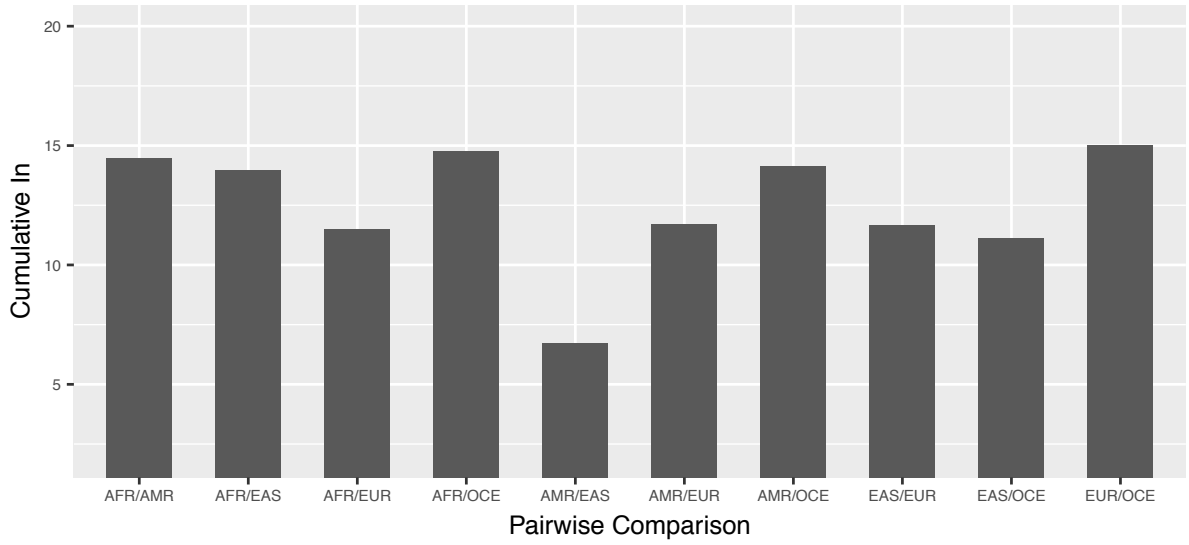


Figure 3. Cumulative I_n values for pairwise population comparisons using 67 biogeographic ancestry SNPs in the custom enrichment panel. AFR: African, AMR: Native America, EAS, East Asian, EUR: European, OCE: Oceanian

STRUCTURE analysis of the reference population data (with no POPFLAG) produced a pattern consistent with five distinct genetic clusters matching the known origin of the reference population samples (Figure 4). Further results from STRUCTURE HARVESTER indicated K=5 produced optimal results using the reference population groups specified previously (Supplementary File S4) as differentiated by the ancestry-informative SNPs in the custom panel.

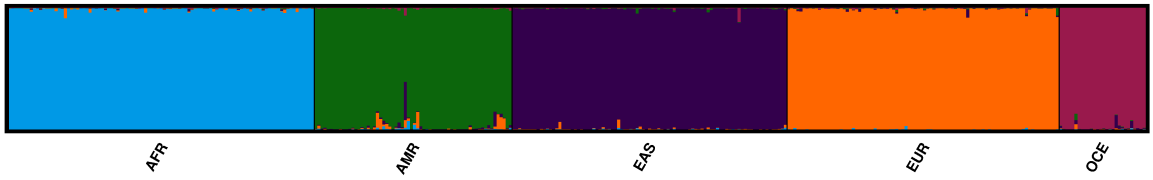


Figure 4. STRUCTURE analysis for reference population data used in this study (optimum clusters as K=5). AFR: African, AMR: Native America, EAS, East Asian, EUR: European, OCE: Oceanian

PCA analyses of the custom enrichment panel against the previously developed Global AIMs Nano set shown in Figure 5 indicate improved separation of all five populations from one another. Further evidence is provided by PC2 vs. PC3, where no sets of points from the reference population groups overlap using the custom SNP enrichment panel.

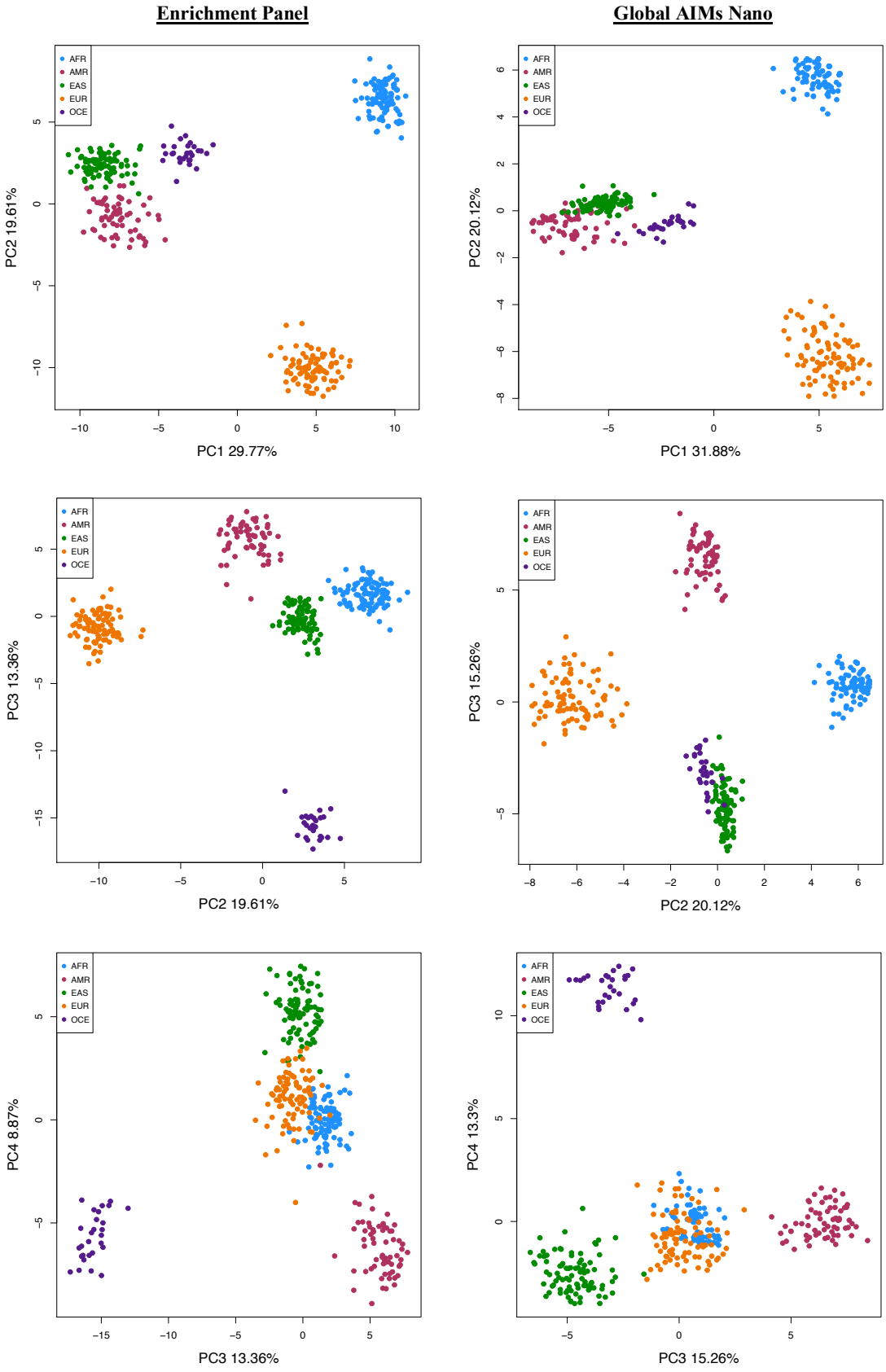


Figure 5. PCA analyses of the 67 biogeographic ancestry SNPs in the custom enrichment panel compared to 31 SNPs in the Global AIMs Nano set for 368 reference population samples (tri-allelic SNPs are not included in analyses). PC: principal component; AFR: African; EUR: European; EAS: East Asian; OCE: Oceanian; AMR: Native American

Sequencing performance of the custom enrichment panel

All 124 SNPs were retrieved from all male samples, and all 89 biogeographic ancestry and phenotype SNPs were recovered from all female samples. With PCR duplicates removed, the average read depth per SNP was 549 ± 278 (Table 7), with a maximum and minimum read depth per SNP of 205-1480 and 51- 623, respectively. The negative control did not recover any reads for any of the SNP markers.

Table 7. Read depth of coverage for 124 SNPs for each of the male samples, and 89 SNPs for females (no Y-chr SNPs). Read depth is reported after PCR duplicate removal (i.e.unique reads only). Average read depth includes haploid markers that show a reduced depth in comparison to autosomal markers.

Sample	Sex	Minimum Read Depth	Maximum Read Depth	Mean Read Depth	No. of Retained Reads
S5	M	66	799	491	499,833
S12	M	72	863	374	492,835
S2	F	532	1290	906	435,675
S3	M	61	327	220	964,251
S6	M	116	1097	735	769,257
S9	F	134	396	240	647,569
S1	M	51	205	146	988,949
S4	F	623	1137	855	912,001
S7	F	593	984	776	711,223
S8	M	102	1480	796	761,713
S10	F	508	844	669	709,279
S11	F	303	678	499	517,490
			Average	549 (±278)	700,840 (±189,733)

The read depth data reported includes coverage for 35 haploid Y-chr markers and two haploid X SNPs (for males) where they are, as expected, noticeably lower than that for autosomal and diploid X SNPs in females (average depth of coverage for Y-chr SNPs is 81-316 versus 420-779 for diploid autosomal SNPs and two X SNPs). The two haploid X SNPs in males gave an average depth of coverage similar to other haploid Y-chr SNPs, whereas the diploid version in females were similar to the autosomal markers (Figure 6). The probes were able to retrieve on average a >500x read depth of coverage for the majority (88%) of biogeographic ancestry and phenotype SNPs, and >200x for the majority (86%) of Y-chr SNPs (Figure 6). Overall, the SNP exhibiting the lowest average depth of coverage resides on the Y-chr (rs2032658), with the highest occurring on chromosome 16 (rs1805006).

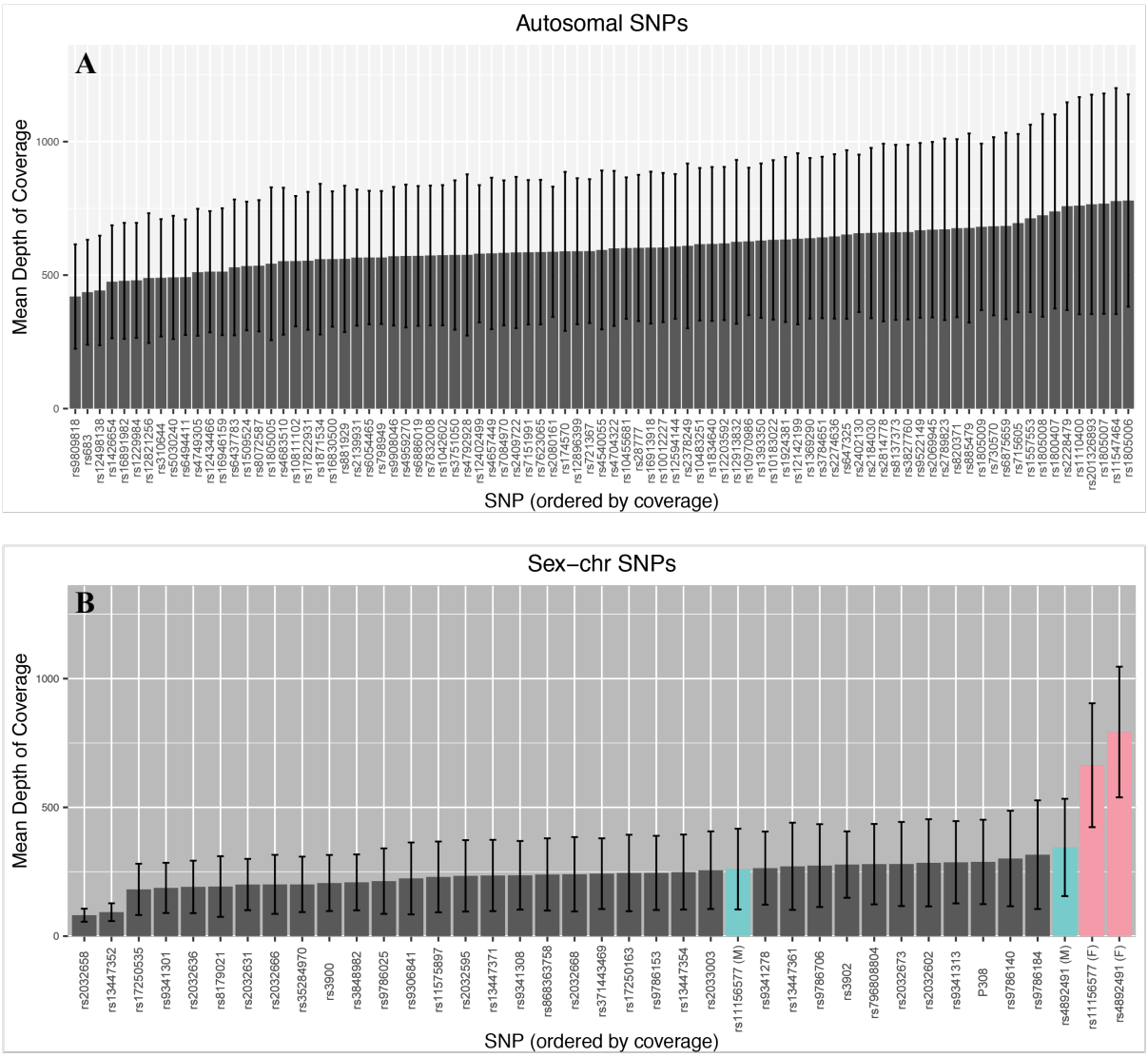


Figure 6. Mean read depth of coverage across 12 samples (six males and six females) for 87 autosomal SNPs (A), and sex chromosome SNPs (two X chromosome SNPs for all 12 samples [rs1115657 and rs4892491], and 35 Y-chr SNPs for six male samples) (B) (in order of increasing coverage). Error bars represent standard deviation. See Supplementary File S6 and S7 for depth of coverage details of each of the SNPs. (M) = Male, blue; (F) = Female, pink.

Biogeographic ancestry assignment

All 67 biogeographic ancestry SNPs were obtained from all twelve test samples. All but three samples had biogeographic ancestry predictions concordant with self-declared ancestry in Snipper (Table 8). Samples which had predictions inconsistent with self-declared ancestry were two individuals with self-declared Native American ancestry (both born in Central America) predicted to have European ancestry, and one individual with self-declared African ancestry (born in North Africa) predicted to have European ancestry. All likelihood ratios were at least 1 billion times more likely one population over any of the other four populations, with the exception of S12 (Table 8).

Table 8. Inferred biogeographic ancestry for 12 samples with self-declared ancestry and their associated likelihood ratios using Snipper. The lowest and highest likelihood ratios are presented to demonstrate the range of values obtained. Remaining likelihood ratios are given in Supplementary Table S8).

Sample	Self-declared ancestry	Region	Snipper Inferred Ancestry	Lowest and Highest Likelihood from Snipper
S5	European	Western Europe	European	1.5E+42 times more likely to be EUR than AMR and 5.9E+66 times more likely to be EUR than OCE
S12	East Asian	South-East Asia	East Asian	654,577,142 times more likely to be EAS than AMR and 4.4E+50 times more likely to be EAS than AFR
S2	Native American	Central America	European	4.6E+38 times more likely to be EUR than AMR and 4.5E+75 times more likely to be EUR than OCE
S3	African	Sub-Saharan Africa	African	9.8E+51 times more likely to be AFR than EUR and 8.2E+79 times likely to be AFR than OCE
S6	European	Western Europe	European	3.4E+57 times more likely to be EUR than AMR and 1.4E+78 times more likely to be EUR than OCE
S9	African	North Africa	European	3.7E+29 times more likely to be EUR than AFR and 1.6E+62 times more likely to be EUR than OCE
S1	European	Western Europe	European	3.5E+51 times more likely to be EUR than AMR and 3.7E+80 times more likely EUR than OCE
S4	European	Western Europe	European	2.6E+36 times more likely to be EUR than AMR and 2E+69 times more likely to be EUR than OCE
S7	European	Western Europe	European	1.6E+40 times more likely to be EUR than AMR and 4.6E+65 times more likely to be EUR than OCE
S8	East Asian	Mainland East Asia	East Asian	3.5E+20 times more likely to be EAS than AMR and 1E+70 times more likely to be EAS than AFR
S10	Native American	Central America	European	3.4E+36 times more likely to be EUR than AMR and 8E+77 times more likely to be EAS than OCE
S11	European	Western Europe	European	3.9E+52 times more likely to be EUR than AMR and 6.4E+75 times more likely to be EUR than OCE

The PCA plot (Figure 7) based on the biogeographic ancestry SNPs for the 12 test samples and 368 population sample genotypes shows clear clustering of the European samples with the EUR reference dataset. Of the two East Asian samples, one clustered on the EAS reference dataset (S8), with the other clustering in the AMR group (S12) despite a Snipper prediction of EAS ancestry with a high likelihood. S10 (self-declared Native American), clustered with the EUR reference population, supported by the Snipper prediction, whereas S2 (self-declared Native American) sits outside the EUR population in the direction of the AMR group. S9 (self-declared African ancestry) also sits outside the EUR population in the direction of the AFR group.

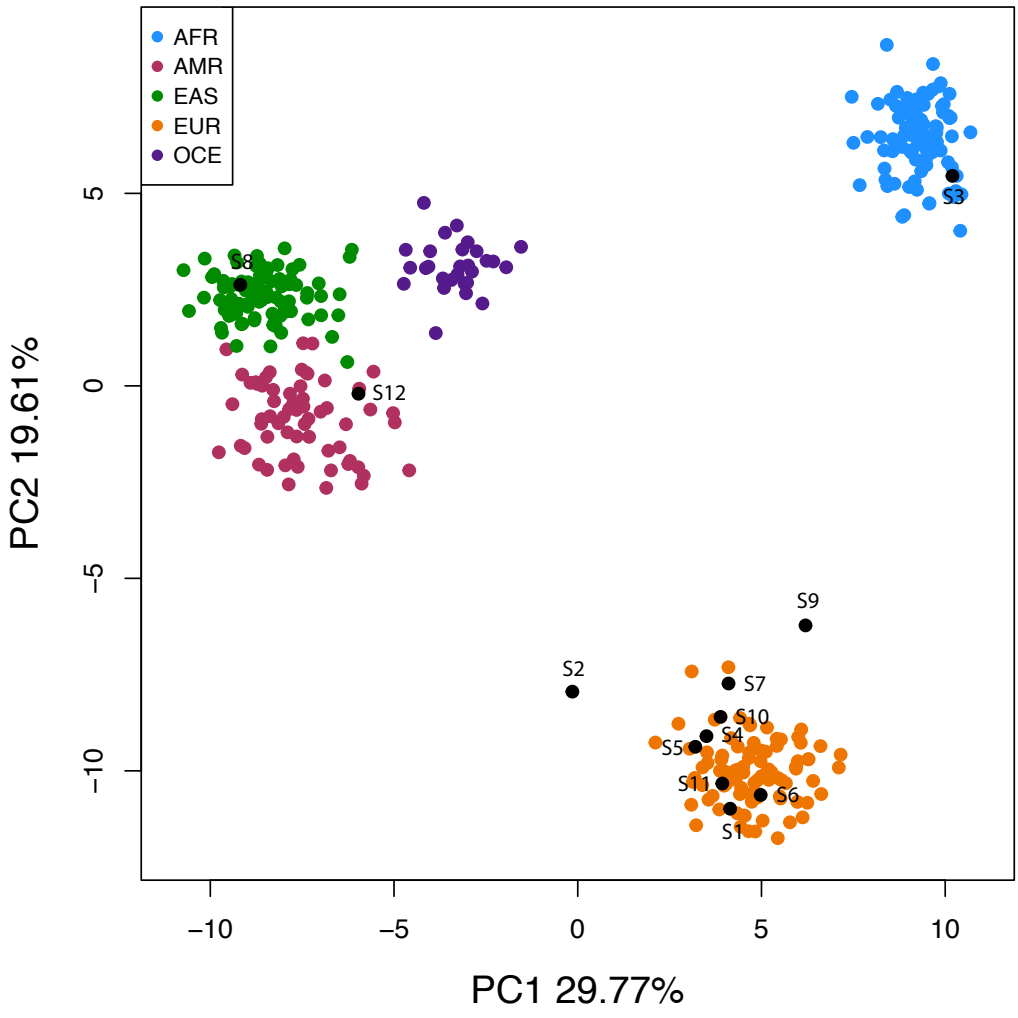


Figure 7. PCA plot of twelve test samples with known (self-declared) ancestry (black circles) and 368 population genotypes representing AFR (blue), EUR (orange), AMR (magenta), EAS (green) and OCE (purple) populations from the 1000 Genomes or HGDP-CEPH datasets. Remaining components are plotted in Supplementary File S9.

Further analysis in STRUCTURE against the same 368 reference population samples show the varying ancestry components in the self-declared samples and was consistent with results

from Snipper and PCA. The admixture bar plot visualised in CLUMPAK shows the estimated cluster membership coefficients produced from STRUCTURE analysis for each of the twelve samples (Figure 8), with further details provided in Table 9. Three samples were shown to have the major ancestry component inconsistent with self-declared ancestry, namely two Native American samples both predicted to have 64.7% and 89.1% EUR ancestry membership, with the AMR component as 29.7% and 9.4%. One North African sample was also shown to have 70.8% EUR ancestry membership, with 26.5% membership to AFR. The remaining nine samples produced a major ancestry component consistent with self-declared ancestry.

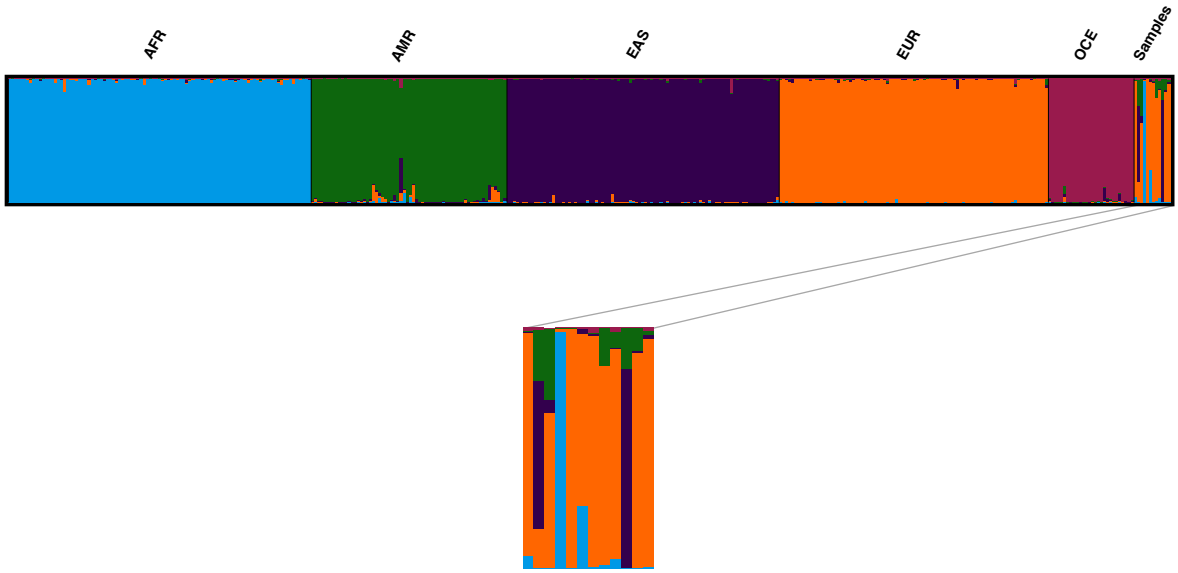


Figure 8. Visualised STRUCTURE analysis of biogeographic ancestry components in twelve test samples with self-declared ancestry. Optimum cluster was K=5. Samples are ordered left to right according to Table 8.

Table 9. Population membership proportions for twelve test samples with self-declared biogeographic ancestry. Grey shading indicates major ancestry component consistent with self-declared ancestry. Red shading indicates a major ancestry component inconsistent with self-declared ancestry.

Sample	Self-declared ancestry	Region	% AFR	% AMR	% EAS	% EUR	% OCE
S5	European	Western Europe	5.1	0.3	0.3	92.8	1.5
S12	East Asian	South-East Asia	0.7	17.8	61.9	16.4	3.2
S2	Native American	Central America	0.2	29.7	4.8	64.7	0.4
S3	African	Sub-Saharan Africa	98.6	0.3	0.1	1	0.1
S6	European	Western Europe	0.1	0.1	0.2	99.5	0.1
S9	African	North Africa	26.5	0.1	2.0	70.8	0.6
S1	European	Western Europe	0.8	0.3	0.6	95.9	2.4
S4	European	Western Europe	2.1	14.5	0.1	83.2	0.1
S7	European	Western Europe	3.5	7.0	0.5	87.2	1.9
S8	East Asian	Mainland East Asia	0.7	16.8	82.0	0.1	0.3
S10	Native American	Central America	0.5	9.4	0.9	89.1	0.1
S11	European	Western Europe	0.7	1.3	1.6	94.8	1.6

Three samples showing major ancestry components inconsistent with self-declared ancestry were plotted on a PCA (Figure 9A) along with four admixed population groups originating from relevant regions of the self-declared samples in order to better assess the possibility of ancestry admixture influencing the results (Table 10).

Table 10. Population groups used as admixed test populations set for testing ancestry admixture using ancestry informative SNPs in the custom enrichment panel. Sample details and genotypes can be found in Supplementary File S10.

Population	N	Data Source	Description
ALG	29	HGDP	Mozabite from Algeria
ASW	61	1000 Genomes	Americans of African ancestry in SW USA
MXL	64	1000 Genomes	Mexican ancestry from Los Angeles
PUR	55	1000 Genomes	Puerto Ricans from Puerto Rico

Further analysis in STRUCTURE of the three samples and four admixed population groups against the same 368 reference population dataset show that the ancestry membership proportions of the study samples fall within the variation for the relevant admixed population groups (i.e. MXL and PUR for Native American samples S2 and S10 and ALG and ASW for S9). The admixture bar plot visualised in CLUMPAK shows the estimated cluster membership coefficients produced from the three test samples and four admixed population groups (Figure 9B). STRUCTURE results were broadly consistent with PCA and indicate ancestry admixture in the three test samples.

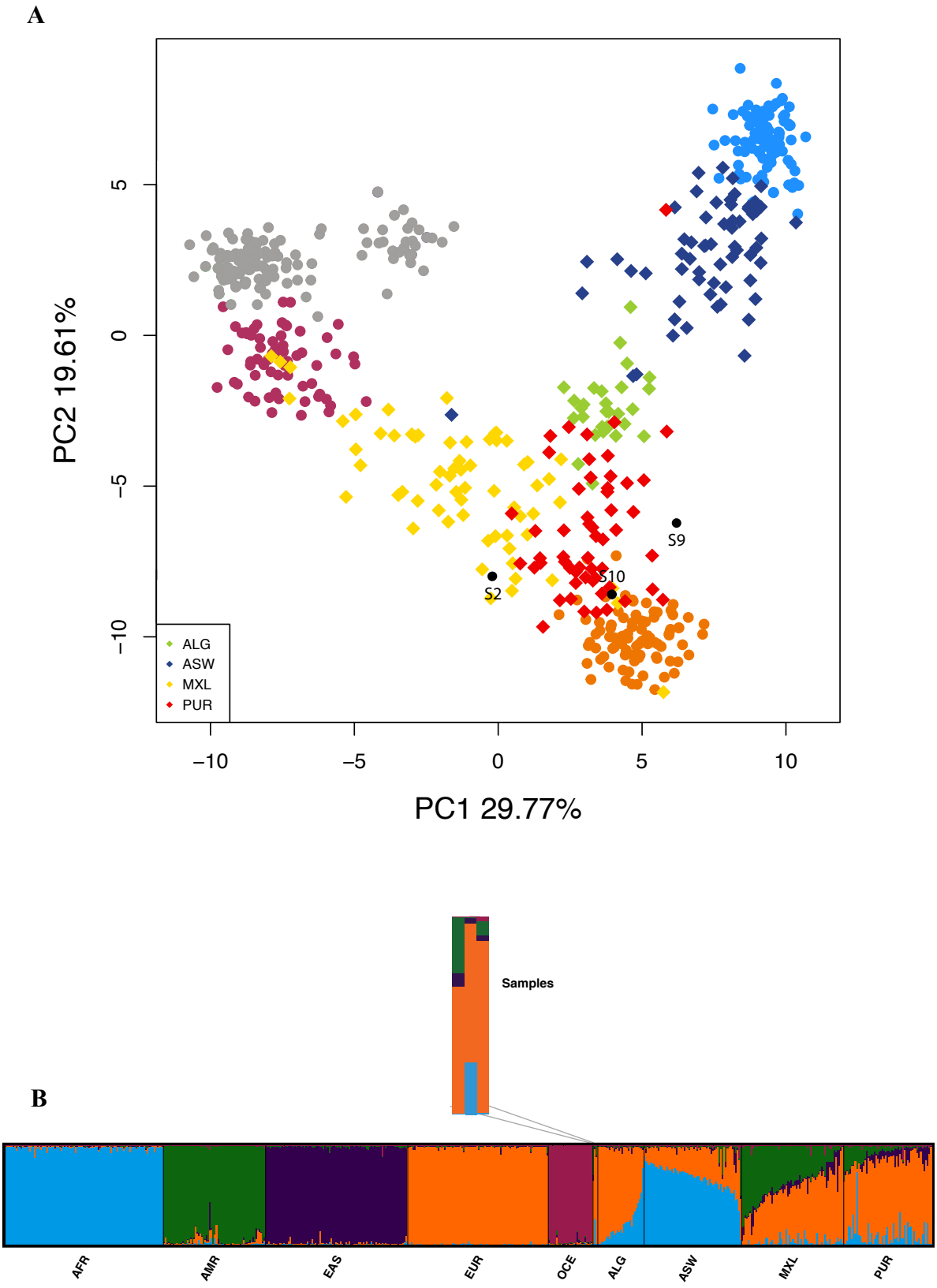


Figure 9. Ancestry analysis of three samples with major ancestry components inconsistent with self-declared ancestry, and four admixed populations (ALG. ASW, MXL, PUR) against five reference populations. (A) PCA plot of samples (black circles), two admixed AFR population groups (ALG: green, ASW: dark blue), two AMR admixed population groups (MXL: yellow, PUR: red), and 368 reference population genotypes. OCE and EAS clusters are greyed out for visual aid. (B) STRUCTURE analysis of three samples (in order of S2, S9, S10), and admixed test populations ALG, ASW, MXL and PUR.

Phenotype prediction

All 23 phenotype SNPs were obtained from all 12 test samples. The HIrisPlex Eye and Hair Colour DNA Phenotyping Webtool correctly predicted eye colour for reported brown and blue eye colours, however incorrectly predicted the intermediate eye colour samples as either blue (S7) or brown (S11) (Table 11). Overall, eye colour was predicted accurately for ten (83%) samples. For hair colour, using the combined 'highest p-value approach' and step-wise model incorporating light or dark shade probability thresholds, predictions were consistent with reported hair colour for eleven samples (92%). One sample (S11) reported as blond, was incorrectly predicted as having brown hair with a light shade (Table 11).

Table 11. Inferred eye colour, most probable hair colour, and associated probabilities for twelve samples with known hair and eye colour using the HIrisPlex SNPs in the custom enrichment panel. D-Brown = Dark Brown; D-blond = Dark Blond.

Sample	SELF-DECLARED		INFERRED PREDICTIONS			
	Eye Colour	Hair Colour	Eye Colour (p-value)	Hair Colour (p-value)	Shade (p-value)	Most Probable Hair Colour
S5	Brown	Brown	Brown (0.767)	Brown (0.475)	Light (0.651)	D-Brown/Black
S12	Brown	Black	Brown (0.998)	Black (0.899)	Dark (0.997)	Black
S2	Brown	Brown	Brown (0.982)	Brown (0.743)	Light (.723)	Brown/D-Brown
S3	Brown	Black	Brown (0.997)	Black (0.679)	Dark (0.997)	Black
S6	Blue	Blond	Blue (0.949)	Blond (0.646)	Light (0.995)	Blond/D-Blond
S9	Brown	Black	Brown (0.991)	Brown (0.498)	Dark (0.905)	D-Brown/Black
S1	Blue	Red	Blue (0.926)	Red (0.846)	Light (0.99)	Red
S4	Blue	Blond	Blue (0.97)	Blond (0.705)	Light (0.996)	Blond
S7	Intermediate	Red	Blue (0.926)	Red (0.4)	Light (0.99)	Red
S8	Brown	Black	Brown (0.957)	Black (0.843)	Dark (0.983)	Black
S10	Brown	Brown	Brown (0.836)	Brown (0.456)	Light (0.583)	D-Brown/Black
S11	Intermediate	Blond	Brown (0.986)	Brown (0.55)	Light (0.849)	Brown

Y-chromosome haplogroup prediction and sex determination

All 35 Y-chr SNPs were recovered from all male test samples. No Y-chr SNPs were called for any of the female samples. Based on the presence versus absence of Y-chr SNPs all twelve samples were predicted accurately as male or female. All Y-chr SNPs made phylogenetic sense (i.e. no conflicting haplogroup assignments), and the inferred Y-haplogroup was broadly consistent with self-declared ancestry for all but one sample from Central America (S2) which carried a Y-chr haplogroup of European origin (Table 12).

Table 12. Y-chr haplogroup results for 12 test samples with known ancestry and associated continental affiliations. 'NA' denotes samples for which a Y-chr haplogroup was not assigned because the donor was female.

Sample	Sex	Self-declared ancestry	Region	Inferred Y-chr haplogroup	Continental Affiliation
S5	M	European	Western Europe	R-M412	Western Europe
S12	M	East Asian	South-East Asia	O-P203	East/South-East Asia
S2	M	Native American	Central America	R-M412	Western Europe
S3	M	African	Sub-Saharan Africa	E-M96	Africa/Europe
S6	M	European	Western Europe	E-V36	Europe
S9	M	African	North Africa	J-M267	North Africa/ Middle East
S1	F	European	Western Europe	NA	NA
S4	F	European	Western Europe	NA	NA
S7	F	European	Western Europe	NA	NA
S8	F	East Asian	Mainland East Asia	NA	NA
S10	F	Native American	Central America	NA	NA
S11	F	European	Western Europe	NA	NA

Discussion

Recent MPS approaches to SNP typing of samples for forensic investigation have shown promise for generating data for hundreds of markers in a single assay (Meiklejohn & Robertson 2017; Xavier & Parson 2017). However, these approaches can present difficulties associated with PCR-based enrichment and customisability of genetic markers targeted (Gettings *et al.* 2015; Elwick *et al.* 2018). We aimed to address these issues by developing a 124-SNP custom hybridisation enrichment panel to infer biogeographic ancestry, hair and eye colour, and Y-chromosome lineage information from a set of modern test samples with a range of self-declared biogeographic ancestry, hair and eye colour, and sex. The use of individually mixed probes also allows the user to customise the panel to suit specific forensic questions on a case-by-case basis. This study demonstrates the feasibility of using this approach for forensic case samples where specific questions regarding biogeographic ancestry and phenotype are of importance.

Analysis of five unadmixed global reference population datasets showed that the 67 ancestry-informative SNPs in the panel were able to accurately differentiate between populations at a continental scale. Cumulative population-specific Divergence (PSD) values indicated virtually equal power for African (AFR) and Oceanian (OCE) populations with a higher than average PSD, while European (EUR) showed slightly lower than average. Both Native American (AMR) and East Asian (EAS) populations showed noticeably lower than average

PSD values. This trend for AMR and EAS population groups has been demonstrated in previous ancestry-informative panels, even for a larger number of markers (Eduardoff *et al.* 2016). This is due to the reduced availability of informative markers divergent for these populations versus population groups like AFR where informative SNPs show higher divergence powers. The lower PSD values can potentially be addressed by adding a small number of extra ancestry SNPs informative for those populations to the panel to achieve a more balanced PSD if required. Pairwise comparisons between population groups showed a reduced power for the SNPs to differentiate between EAS and AMR, and EAS and OCE, which can be explained by their much more recent shared common ancestry and the similar allele frequencies for many SNPs in these populations (de la Puente *et al.* 2017). Despite this, the custom panel was able to differentiate between populations to a higher degree than previous SNaPshot SNP typing performed in the laboratory using the Global AIMs Nano set. The inclusion of extra SNPs specifically informative for pairwise differentiation of EAS populations from AMR and OCE will improve the ability of the panel to differentiate between these populations.

Given the ability of the panel to be fully customisable, by adding or removing probes for individual SNPs or groups of SNPs, analyses can be tailored to suit specific questions of forensic testing. For example, phenotypic SNP data might not be required and can be omitted from the enrichment probe pool, or population differentiation may be required only for a pairwise or three-way comparison (i.e. differentiating between AFR and EUR ancestry only) (Phillips *et al.* 2007; Phillips *et al.* 2009). In these cases, the careful selection of specific markers informative for those groups can be targeted to suit each investigation on a case-by-case basis. PSD and pairwise comparisons should be conducted during the marker selection process for these situations to ensure adequate differentiation power of the population groups in question. While this is true for the biogeographic ancestry informative SNPs, this is not a concern for the phenotypic SNPs and no such calculations are required for the Y-chr lineage SNPs in the panel. The customisability of the panel is desirable when attempting to tailor forensic analyses and it addresses the ‘big data’ concern of MPS for forensic identification by reducing the amount of data generated whilst ensuring the requirements for testing are met (Scudder *et al.* 2018). The use of tailored ancestry SNP sets for specific population comparisons will theoretically reduce the error associated with the interpretation of ancestry markers (Phillips *et al.* 2009).

A combination of three ancestry classification methods (Snipper, STRUCTURE and PCA) were used to infer biogeographic ancestry in this study, given the limitations of Snipper to detect ancestry admixture (Cheung *et al.* 2018a). The biogeographic ancestries of all European and East Asian declared individuals were correctly inferred from the SNPs using Snipper, STRUCTURE and PCA, however samples with Native American declared ancestry from Central America and one individual with African declared ancestry had major ancestry components inconsistent with declared information, clustering strongly to European ancestry. Most modern Central and South American populations have varying degrees of ancestry admixture from European, Native American and African ancestry (Galanter *et al.* 2012; Phillips *et al.* 2014; Homburger *et al.* 2015; de la Puente *et al.* 2017). Our results for the two self-declared Native American samples therefore aligned with previous analysis of individuals from this geographic region and further demonstrates the complexity of assigning ancestry to admixed Central and South American individuals and populations. North African populations have also demonstrated gene flow and shared ancestry from the neighbouring European continent and the Near East, revealing admixture between African and European ancestry and are differentiated from other Sub-Saharan populations in previous studies (Henn *et al.* 2012; Jin *et al.* 2018). Furthermore, North African populations can more closely resemble European populations than sub-Saharan African using ancestry informative SNPs commonly used in forensic ancestry panels (Phillips 2013), also evident in the analyses of SNP data from an Algerian (ALG) population of the HDGP-CEPH dataset in this study. Comparison of the North African sample against the ALG population indicated consistency between the estimated co-ancestry components using STRUCTURE and can therefore be considered an admixed individual of AFR and EUR ancestry. It is important to note that for testing self-declared individuals who have admixed ancestry, not all contributing ancestries may be declared and must be taken into account when reporting genetic ancestry of test samples. Recent guidelines proposed suggest that an ancestry component should be reported for a sample if it contributes to more than 20% of the overall ancestry profile (Jin *et al.* 2018). However, empirical thresholds selected would depend on the specific panel used, the classification methods used to analyse the data, as well as the reference population genotypes that are available for comparison. This would require deliberation in each laboratory depending on what methods are in use.

Our results support previous concerns around ancestry determination for forensic testing, particularly when attempting to categorise an individual from a modern (and therefore potentially admixed) population into one of five ancestral reference population groups

(Cheung *et al*. 2018b; Jin *et al*. 2018). For example, the two Native American declared samples from Central America in our study display high levels of European admixture and therefore do not group strongly into the unadmixed Native American reference population dataset. The differentiation of admixed populations and those on continental borders where no geographic boundaries exist can complicate analyses due to increased gene flow, especially in the modern age where movement between continents is common. As a potential approach, it may be beneficial to include admixed populations as well as 'unadmixed' reference population groups in order to compare and better assess ancestry in unknown samples, rather than evaluating samples against unadmixed reference populations alone. On the whole, given this information the predictions in this study that were inconsistent with self-declared ancestry were not considered as erroneous. The panel had correctly identified ancestry components of these samples that align with previous studies of relevant admixed populations (de la Puente *et al.* 2017; Henn *et al.* 2012). Additionally, these samples show ancestry components consistent with the geography of the regions where the self-declared individuals originated from (and likely admixture due to shared demographic history).

Samples selected in this study were chosen to represent all hair and eye colour categories to test whether the inclusion of the phenotype SNPs with the ancestry and Y-chr SNPs using a hybridisation enrichment technology gave results that were consistent with known phenotype. Brown and blue eye colours were predicted accurately in all cases, however intermediate eye colours remained problematic to predict, giving an overall 83% prediction accuracy of the SNPs to infer eye colour (Walsh *et al.* 2014). Interestingly, when excluding the intermediate eye colour category (sometimes explored due to the potential inaccuracies in predicting intermediate eye colour against observed eye colour)(Walsh *et al.* 2014), the prediction accuracy increases to 92% when grouping individuals into 'brown' and 'not brown' eye colour categories. Given that pigmentation in eye colour is a complex trait which can be subjective to report (Sulem *et al.* 2008), and that intermediate eye colour has demonstrated a lower prediction accuracy than other eye colours in previous studies (Ruiz *et al.* 2013; Freire-Aradas *et al.* 2014; Walsh *et al.* 2014; Hussing *et al.* 2015), this result is not unexpected. For hair colour, a 92% prediction accuracy was achieved across the twelve samples, where one sample with reported blond hair colour was predicted as brown. Red, black and brown hair phenotypes were predicted correctly in all cases. Again, previous studies have documented inaccuracies with predicting hair colour phenotypes (down to a 73% prediction accuracy on average), particularly with blond and brown categories (Walsh *et al.* 2014; Hussing *et al.* 2015). For both hair and eye colour, the prediction accuracy shown in this study is consistent

with previous error rates established in earlier studies of the HIrisPlex SNP panel (Walsh *et al.* 2013; Walsh *et al.* 2014). Since the design and execution of our panel, a revised HIrisPlex panel has been published, termed the 'HIrisPlex-S' assay which now includes 17 extra SNPs in pigmentation genes which can help infer skin colour (Chaitanya *et al.* 2018). As an additional consideration, these SNPs could easily be implemented into the custom enrichment panel as a further intelligence tool. Nonetheless, this study has demonstrated the successful use of the HIrisPlex panel in a hybridisation enrichment approach for forensic analysis and may help to further support ancestry estimations when used in conjunction with the ancestry informative SNPs in the custom panel. Currently, the HIrisPlex model includes test data only from European populations (Walsh *et al.* 2014). Understanding how different populations may influence the prediction model and therefore the success rate could be improved by including reference samples from multiple non-European populations.

The Y-chr SNPs in the custom panel were able to predict Y-chr haplogroups for all samples with no conflicting haplogroup classifications. No Y-chr SNPs were recovered from any of the female samples, highlighting the potential for this method as an indication of sex (based on the presence versus absence of Y-chr markers). For all six male samples, Y-chr haplogroup classifications and their associated most likely geographic affiliation were consistent with reported self-declared ancestry with the exception of one self-declared Native American sample (with a European biogeographic ancestry prediction), carrying a European Y-chr haplogroup (R-M412). Given the complex demographic history of Central and South American populations discussed above, this result is not unexpected and represents the European influence on modern American ancestry. The distinction between sub-haplogroups of widespread haplogroups has demonstrated utility in this study by identifying two different E haplogroups that have contrasting geographic coverage (Cruciani *et al.* 2007; Valverde *et al.* 2013; van Oven *et al.* 2013). Further resolution of the E haplogroup allowed for more specific geographical affiliation of patrilineal ancestry, and was able to distinguish between two samples of African and European ancestry both carrying E lineages that have differing geographical coverage (van Oven *et al.* 2013). Haplogroup J-M267 detected in the North African sample which revealed predominantly European autosomal ancestry, is common in the Middle East and North Africa and was therefore consistent with known information (Semino *et al.* 2004; van Oven *et al.* 2013). The panel has the capability of determining informative Y-chr haplogroups and sub-haplogroups and can be considered a suitable tool for exploring the paternal lineage of male samples. In each case, Y-chr haplogroup assignments complemented the results of the biogeographic ancestry analysis and was able to further

resolve ancestry components, especially for male samples showing evidence of ancestry admixture in autosomal SNP analysis.

Conclusion

The custom enrichment panel provides a new avenue for the simultaneous genetic analysis of multiple marker types for forensic intelligence (biogeographic ancestry, phenotype and paternal lineage), with a novel technical approach that allows the possibility of using customisable SNP marker sets for hybridisation enrichment prior to MPS. The panel can distinguish biogeographic ancestry of a sample between five major continental populations and detect admixed individuals. The inclusion of Y-chr SNPs demonstrated fine-resolution sub-typing of ubiquitous haplogroups in order to assign more specific geographical affiliation of paternal lineages and was able to support findings from biogeographic ancestry analysis. Hair and eye colour predictions from a range of hair colours using SNPs of the HIrisPlex phenotyping tool produced predictions that match well with previously established success rates. Thus, the panel provided genetic information for a range of modern samples with known ancestry, sex and phenotype consistent with self-declared ancestry and demographic histories of the regions they originated from. The panel demonstrates accurate inference of ancestry, sex, paternal lineage and hair and eye colour, all of which provide valuable intelligence information for questions of forensic testing involving cases of missing persons and historical human remains.

Competing Interests

The authors declare no competing interests

Acknowledgements

We thank the volunteers for generously donating a DNA sample to assist in the project. The research was supported by an Australian Research Council (ARC) Future Fellowship (FT10010008), ARC Discovery Project (DP150101664) and ARC LIEF Project (LE160100154) to JJA.

References

Aird, D., Ross, M.G., Chen, W.-S., Danielsson, M., Fennell, T., Russ, C., Jaffe, D.B., Nusbaum, C. & Gnirke, A. 2011. Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries, *Genome Biology*, 12, R18.

Bergström, A., Nagle, N., Chen, Y., McCarthy, S., Pollard, Martin O., Ayub, Q., Wilcox, S., Wilcox, L., van Oorschot, Roland A., McAllister, P., et al. 2016. Deep Roots for Aboriginal Australian Y Chromosomes, *Curr Biol*, 26, 809-13.

Bose, N., Carlberg, K., Sensabaugh, G., Erlich, H. & Calloway, C. 2018. Target capture enrichment of nuclear SNP markers for massively parallel sequencing of degraded and mixed samples, *Forensic Sci Int Genet*, 34, 186-96.

Budowle, B. & van Daal, A. 2008. Forensically relevant SNP classes, *Biotechniques*, 44, 603-8, 10.

Chaitanya, L., Breslin, K., Zuñiga, S., Wirken, L., Pośpiech, E., Kukla-Bartoszek, M., Sijen, T., Knijff, P.d., Liu, F., Branicki, W., et al. 2018. The HIrisPlex-S system for eye, hair and skin colour prediction from DNA: Introduction and forensic developmental validation, *Forensic Sci Int Genet*, 35, 123-35.

Cheung, E.Y.Y., Gahan, M.E. & McNevin, D. 2018a. Prediction of biogeographical ancestry in admixed individuals, *Forensic Sci Int Genet*, 36, 104-11.

Cheung, E.Y.Y., Gahan, M.E. & McNevin, D. 2018b. Predictive DNA analysis for biogeographical ancestry, *Aust J Forensic Sci*, 1-8.

Cruciani, F., La Fratta, R., Trombetta, B., Santolamazza, P., Sellitto, D., Colomb, E.B., Dugoujon, J.-M., Crivellaro, F., Benincasa, T., Pascone, R., et al. 2007. Tracing Past Human Male Movements in Northern/Eastern Africa and Western Eurasia: New Clues from Y-Chromosomal Haplogroups E-M78 and J-M12, *Mol Biol Evol*, 24, 1300-11.

Dabney, J. & Meyer, M. 2012. Length and GC-biases during sequencing library amplification: A comparison of various polymerase-buffer systems with ancient and modern DNA sequencing libraries, *Biotechniques*, 52, 87-94.

Daca-Roszak, P., Pfeifer, A., Żebracka-Gala, J., Jarząb, B., Witt, M. & Ziętkiewicz, E. 2016. EurEAs_Gplex—A new SNaPshot assay for continental population discrimination and gender identification, *Forensic Sci Int Genet*, 20, 89-100.

Daya, M., van der Merwe, L., Galal, U., Möller, M., Salie, M., Chimusa, E.R., Galanter, J.M., van Helden, P.D., Henn, B.M., Gignoux, C.R., et al. 2013. A Panel of Ancestry Informative Markers for the Complex Five-Way Admixed South African Coloured Population, *PLoS One*, 8, e82224.

de la Puente, M., Phillips, C., Santos, C., Fondevila, M., Carracedo, Á. & Lareu, M.V. 2017. Evaluation of the Qiagen 140-SNP forensic identification multiplex for massively parallel sequencing, *Forensic Sci Int Genet*, 28, 35-43.

de la Puente, M., Santos, C., Fondevila, M., Manzo, L., Carracedo, A., Lareu, M.V. & Phillips, C. 2016. The Global AIMs Nano set: A 31-plex SNaPshot assay of ancestry-informative SNPs, *Forensic Sci Int Genet*, 22, 81-8.

Earl, D.A. & vonHoldt, B.M. 2012. STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method, *Conserv Genet Resour*, 4, 359-61.

Eduardoff, M., Gross, T.E., Santos, C., de la Puente, M., Ballard, D., Strobl, C., Børsting, C., Morling, N., Fusco, L., Hussing, C., et al. 2016. Inter-laboratory evaluation of the EUROFORGEN Global ancestry-informative SNP panel by massively parallel sequencing using the Ion PGM™, *Forensic Sci Int Genet*, 23, 178-89.

Elwick, K., Zeng, X., King, J., Budowle, B. & Hughes-Stamm, S. 2018. Comparative tolerance of two massively parallel sequencing systems to common PCR inhibitors, *Int J Legal Med*, 132, 983-95.

Freire-Aradas, A., Ruiz, Y., Phillips, C., Maronas, O., Sochtig, J., Tato, A.G., Dios, J.A., de Cal, M.C., Silbiger, V.N., Luchessi, A.D., et al. 2014. Exploring iris colour prediction and ancestry inference in admixed populations of South America, *Forensic Sci Int Genet*, 13, 3-9.

Galanter, J.M., Fernandez-Lopez, J.C., Gignoux, C.R., Barnholtz-Sloan, J., Fernandez-Rozadilla, C., Via, M., Hidalgo-Miranda, A., Contreras, A.V., Figueroa, L.U., Raska, P., et al. 2012. Development of a Panel of Genome-Wide Ancestry Informative Markers to Study Admixture Throughout the Americas, *PLoS Genet*, 8, e1002554.

Gettings, K.B., Kiesler, K.M. & Vallone, P.M. 2015. Performance of a next generation sequencing SNP assay on degraded DNA, *Forensic Sci Int Genet*, 19, 1-9.

Gonzalez, J.R., Armengol, L., Sole, X., Guino, E., Mercader, J.M., Estivill, X. & Moreno, V. 2007. SNPassoc: an R package to perform whole genome association studies, *Bioinformatics*, 23, 644-5.

Henn, B.M., Botigué, L.R., Gravel, S., Wang, W., Brisbin, A., Byrnes, J.K., Fadhlaoui-Zid, K., Zalloua, P.A., Moreno-Estrada, A., Bertranpetit, J., et al. 2012. Genomic Ancestry of North Africans Supports Back-to-Africa Migrations, *PLoS Genet*, 8, e1002397.

Homburger, J.R., Moreno-Estrada, A., Gignoux, C.R., Nelson, D., Sanchez, E., Ortiz-Tello, P., Pons-Estel, B.A., Acevedo-Vasquez, E., Miranda, P., Langefeld, C.D., et al. 2015. Genomic Insights into the Ancestry and Demographic History of South America, *PLoS Genet*, 11, e1005602.

Hudjashov, G., Kivisild, T., Underhill, P.A., Endicott, P., Sanchez, J.J., Lin, A.A., Shen, P., Oefner, P., Renfrew, C., Villems, R., et al. 2007. Revealing the prehistoric settlement of Australia by Y chromosome and mtDNA analysis, *Proc Natl Acad Sci*, 104, 8726.

Hussing, C., Børsting, C., Mogensen, H.S. & Morling, N. 2015. Testing of the Illumina® ForenSeq™ kit, *Forensic Sci Int Genet Supp Series*, 5, e449-e50.

Jin, S., Chase, M., Henry, M., Alderson, G., Morrow, J.M., Malik, S., Ballard, D., McGory, J., Fernandopulle, N., Millman, J., et al. 2018. Implementing a biogeographic ancestry inference service for forensic casework, *Electrophoresis*, 39, 2757-65.

Karafet, T.M., Hallmark, B., Cox, M.P., Sudoyo, H., Downey, S., Lansing, J.S. & Hammer, M.F. 2010. Major east-west division underlies Y chromosome stratification across Indonesia, *Mol Biol Evol*, 27, 1833-44.

Karafet, T.M., Mendez, F.L., Meilerman, M.B., Underhill, P.A., Zegura, S.L. & Hammer, M.F. 2008. New binary polymorphisms reshape and increase resolution of the human Y chromosomal haplogroup tree, *Genome Res*, 18, 830-8.

Karafet, T.M., Mendez, F.L., Sudoyo, H., Lansing, J.S. & Hammer, M.F. 2015. Improved phylogenetic resolution and rapid diversification of Y-chromosome haplogroup K-M526 in Southeast Asia, *Eur J Hum Genet*, 23, 369-73.

Kosoy, R., Nassir, R., Tian, C., White Phoebe, A., Butler Lesley, M., Silva, G., Kittles, R., Alarcon-Riquelme Marta, E., Gregersen Peter, K., Belmont John, W., et al. 2008. Ancestry informative marker sets for determining continental origin and admixture proportions in common populations in America, *Human Mutat*, 30, 69-78.

Lao, O., Vallone, P.M., Coble, M.D., Diegoli, T.M., van Oven, M., van der Gaag, K.J., Pijpe, J., de Knijff, P. & Kayser, M. 2010. Evaluating Self-declared Ancestry of U.S. Americans with Autosomal, Y-chromosomal and Mitochondrial DNA, *Human Mutat*, 31, e1875-e93.

Li, H. & Durbin, R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform, *Bioinformatics*, 25, 1754-60.

Lindgreen, S. 2012. AdapterRemoval: easy cleaning of next-generation sequencing reads, *BMC Research Notes*, 5, 337.

Meiklejohn, K.A. & Robertson, J.M. 2017. Evaluation of the Precision ID Identity Panel for the Ion Torrent™ PGM™ sequencer, *Forensic Sci Int Genet*, 31, 48-56.

Meyer, M. & Kircher, M. 2010. Illumina sequencing library preparation for highly multiplexed target capture and sequencing, *Cold Spring Harb Protoc*, 2010, pdb.prot5448.

Musgrave-Brown, E., Ballard, D., Balogh, K., Bender, K., Berger, B., Bogus, M., Børsting, C., Brion, M., Fondevila, M., Harrison, C., et al. 2007. Forensic validation of the SNPforID 52-plex assay, *Forensic Sci Int Genet*, 1, 186-90.

Nagle, N., Ballantyne, K.N., van Oven, M., Tyler-Smith, C., Xue, Y., Taylor, D., Wilcox, S., Wilcox, L., Turkalov, R., van Oorschot, R.A.H., et al. 2015. Antiquity and diversity of aboriginal Australian Y-chromosomes, *Am J Phys Anthropol*, 159, 367-81.

Naitoh, S., Kasahara-Nonaka, I., Minaguchi, K. & Nambiar, P. 2013. Assignment of Y-chromosomal SNPs found in Japanese population to Y-chromosomal haplogroup tree, *J Hum Genet*, 58, 195.

Park, M.J., Lee, H.Y., Kim, N.Y., Lee, E.Y., Yang, W.I. & Shin, K.J. 2013. Y-SNP miniplexes for East Asian Y-chromosomal haplogroup determination in degraded DNA, *Forensic Sci Int Genet*, 7, 75-81.

Phillips, C. 2013, 'Ancestry Informative Markers', in JA Siegel, PJ Saukko & MM Houck (eds), *Encyclopedia of Forensic Sciences (Second Edition)*, Academic Press, Waltham, pp. 323-31.

Phillips, C. 2015. Forensic genetic analysis of bio-geographical ancestry, *Forensic Sci Int Genet*, 18, 49-65.

Phillips, C., Parson, W., Lundsberg, B., Santos, C., Freire-Aradas, A., Torres, M., Eduardoff, M., Børsting, C., Johansen, P., Fondevila, M., et al. 2014. Building a forensic ancestry panel from the ground up: The EUROFORGEN Global AIM-SNP set, *Forensic Sci Int Genet*, 11, 13-25.

Phillips, C., Prieto, L., Fondevila, M., Salas, A., Gómez-Tato, A., Álvarez-Dios, J., Alonso, A., Blanco-Verea, A., Brión, M., Montesino, M., et al. 2009. Ancestry Analysis in the 11-M Madrid Bomb Attack Investigation, *PLoS One*, 4, e6583.

Phillips, C., Salas, A., Sanchez, J.J., Fondevila, M., Gomez-Tato, A., Alvarez-Dios, J., Calaza, M., de Cal, M.C., Ballard, D., Lareu, M.V., et al. 2007. Inferring ancestral origin using a single multiplex assay of ancestry-informative marker SNPs, *Forensic Sci Int Genet*, 1, 273-80.

Rosenberg, N.A., Li, L.M., Ward, R. & Pritchard, J.K. 2003. Informativeness of Genetic Markers for Inference of Ancestry, *Am J Hum Genet*, 73, 1402-22.

Ruiz, Y., Phillips, C., Gomez-Tato, A., Alvarez-Dios, J., Casares de Cal, M., Cruz, R., Maroñas, O., Söchtig, J., Fondevila, M., Rodriguez-Cid, M.J., et al. 2013. Further development of forensic eye color predictive tests, *Forensic Sci Int: Genet*, 7, 28-40.

Santos, C., Phillips, C., Fondevila, M., Daniel, R., van Oorschot, R.A.H., Burchard, E.G., Schanfield, M.S., Souto, L., Uacyisrael, J., Via, M., et al. 2016. Pacifiplex: an ancestry-informative SNP panel centred on Australia and the Pacific region, *Forensic Sci Int Genet*, 20, 71-80.

Schubert, M., Ermini, L., Der Sarkissian, C., Jonsson, H., Ginolhac, A., Schaefer, R., Martin, M.D., Fernandez, R., Kircher, M., McCue, M., et al. 2014. Characterization of ancient and modern genomes by SNP detection and phylogenomic and metagenomic analysis using PALEOMIX, *Nat Protoc*, 9, 1056-82.

Scudder, N., McNevin, D., Kelty, S.F., Walsh, S.J. & Robertson, J. 2018. Massively parallel sequencing and the emergence of forensic genomics: Defining the policy and legal issues for law enforcement, *Science & Justice*, 58, 153-8.

Semino, O., Magri, C., Benuzzi, G., Lin, A.A., Al-Zahery, N., Battaglia, V., Maccioni, L., Triantaphyllidis, C., Shen, P., Oefner, P.J., et al. 2004. Origin, diffusion, and differentiation of Y-chromosome haplogroups E and J: inferences on the neolithization of Europe and later migratory events in the Mediterranean area, *Am J Hum Genet*, 74, 1023-34.

Sulem, P., Gudbjartsson, D.F., Stacey, S.N., Helgason, A., Rafnar, T., Jakobsdottir, M., Steinberg, S., Gudjonsson, S.A., Palsson, A., Thorleifsson, G., et al. 2008. Two newly identified genetic determinants of pigmentation in Europeans, *Nature Genet*, 40, 835.

Taylor, D., Nagle, N., Ballantyne, K.N., van Oorschot, R.A., Wilcox, S., Henry, J., Turakulov, R. & Mitchell, R.J. 2012. An investigation of admixture in an Australian Aboriginal Y-chromosome STR database, *Forensic Sci Int Genet*, 6, 532-8.

Templeton, J.E.L., Brotherton, P.M., Llamas, B., Soubrier, J., Haak, W., Cooper, A. & Austin, J.J. 2013. DNA capture and next-generation sequencing can recover whole mitochondrial genomes from highly degraded samples for human identification, *Investig Genet*, 4, 26.

Tewhey, R., Nakano, M., Wang, X., Pabón-Peña, C., Novak, B., Giuffre, A., Lin, E., Happe, S., Roberts, D.N., LeProust, E.M., et al. 2009. Enrichment of sequencing targets from the human genome by solution hybridization, *Genome Biology*, 10, R116-R.

Underhill, P.A., Poznik, G.D., Rootsi, S., Järve, M., Lin, A.A., Wang, J., Passarelli, B., Kanbar, J., Myres, N.M., King, R.J., et al. 2014. The phylogenetic and geographic structure of Y-chromosome haplogroup R1a, *Eur J Hum Genet*, 23, 124.

Valverde, L., Köhnemann, S., Cardoso, S., Pfeiffer, H. & de Pancorbo Marian, M. 2013. Improving the analysis of Y-SNP haplogroups by a single highly informative 16 SNP multiplex PCR-minisequencing assay, *Electrophoresis*, 34, 605-12.

van Oven, M., Ralf, A. & Kayser, M. 2011. An efficient multiplex genotyping approach for detecting the major worldwide human Y-chromosome haplogroups, *Int J Legal Med*, 125, 879-85.

van Oven, M., Toscani, K., Tempel, N., Ralf, A. & Kayser, M. 2013. Multiplex genotyping assays for fine-resolution subtyping of the major human Y-chromosome haplogroups E, G, I, J, and R in anthropological, genealogical, and forensic investigations, *Electrophoresis*, 34, 3029-38.

van Oven, M., van den Tempel, N. & Kayser, M. 2012. A multiplex SNP assay for the dissection of human Y-chromosome haplogroup O representing the major paternal lineage in East and Southeast Asia, *J Hum Genet*, 57, 65-9.

Walsh, S., Chaitanya, L., Clarisse, L., Wirken, L., Draus-Barini, J., Kovatsi, L., Maeda, H., Ishikawa, T., Sijen, T., de Knijff, P., et al. 2014. Developmental validation of the HIrisPlex system: DNA-based eye and hair colour prediction for forensic and anthropological usage, *Forensic Sci Int Genet*, 9, 150-61.

Walsh, S., Liu, F., Wollstein, A., Kovatsi, L., Ralf, A., Kosiniak-Kamysz, A., Branicki, W. & Kayser, M. 2013. The HIrisPlex system for simultaneous prediction of hair and eye colour from DNA, *Forensic Sci Int Genet*, 7, 98-115.

Xavier, C. & Parson, W. 2017. Evaluation of the Illumina ForenSeq™ DNA Signature Prep Kit – MPS forensic application for the MiSeq FGx™ benchtop sequencer, *Forensic Sci Int Genet*, 28, 188-94.

Zhong, H., Shi, H., Qi, X.-B., Xiao, C.-J., Jin, L., Ma, R.Z. & Su, B. 2010. Global distribution of Y-chromosome haplogroup C reveals the prehistoric migration routes of African exodus and early settlement in East Asia, *J Hum Genet*, 55, 428.

Supplementary Information

Supplementary Table S1. Bait sequence (120 bp) and GC content for the 67 biogeographic ancestry SNPs included in the hybridisation enrichment panel. rs numbers in bold designate tri-allelic SNPs. Position of the targeted SNPs are shown in red. rs16891982 is also included in the phenotype SNP set. SNPs are listed in the same order as in Table 1.

rs Number	Chr.	Position (GRCh37/hg19)	Bait Sequence	GC content
rs2139931	1	84590527	TCTGGAGATTTAGTTATGGGACTACCTGGCTATAGTCTTGGCTAGGGCGTTAGTAGATCAATGAAA TTTTCTCATCAACACGAAGGGCTCTAAGTGTTAAAGGGAGACAGTGCCCACATC	44.2
rs2814778	1	159174683	GTTTGGTTCAGGCCCGCAGACAGAAGGGCTGGGACGGCTGTCAGCGCCTGTGCTTCCAAGATAAG AGCCAAGGACTAATGAGGGCCATCAGGTTCTGGGCAGGGACAGGGGGCATCAGGA	60.0
rs4657449	1	165465281	TTAATACCCCCTCGGGAGAAAACATAGAAAGTTATGAGCTGAGCTAAGGAAAGATACGTGGATGC AAATCTGACCCTCATTCAACTCTAGAAAATTCAGCTGGAGAAAAAACATGTTATT	38.3
rs12142199	1	1249187	GTCAGCAAAAGCCCGGTCGAAGGCCTTGATGTGCTTGAACTCAAACATGTTCCTCTGCACGAAAG TCTTGCGGATCTTCTGGTTGGTCCAGGGGATGAACAGCTTGTAGTAGTGGTTGGC	51.7
rs12402499	1	101528954	ATTGTTGATTAAGAGAACAGATTTATTGACAGACTTCTGCTTTTGATTTCAAGTATCAGTGTTGATT GACTGTCTGCTGTAAGCATAATAGTGATGGATATAAAGCCACTAGTAATACCC	34.2
rs647325	1	18170886	GCACAGATGGTGCAATTAGAGGCTTGGGCAGATTCGCTGTAGAAGATGATGCCACTGGAGAAAAG GAGCCAGTTCTGCGGGGGGTCTTCTCTATTTTAAAAGCAGCCGGCAGTGTATAAT	49.2
rs2184030	1	206667441	GGGGGTGGTGAAGAGTCCCCAAGCCTGCCCATGCAGGTTTGATGACACGTGTCCATTTTCGAAGA AGCCCTGTTGAAGCTTAGGCATTTTCCACAAATGAGTTGACATCCTCAGGAGGTG	51.7
rs16830500	2	152814129	AAGAACAGGGAAACTAAGTGCTGTGTGCACTGTCACAAACTCTCGAGCTGACCAAGACAATAATT TGACCCAGTATTTCTCAAAGAATTTACATGAAAGTGATTTGGGGAGCGTTTGCCA	41.7
rs3827760	2	109513601	TTGATTGCCTCGAGAAGACTAGCCGAATGCTCAGCTCCACGTACAACTCTGAGAAGGCTGTTGTG AAAACGTGGCGCCACCTCGCCGAGAGCTTCGGCCTGAAGAGGGATGAGATTGGGG	55.8
rs10183022	2	237481969	CTAACTTGCAGGCTCAGCGTAAGCGGGGCTGGCCTGGTGGAAGAGGTGTTATCTGCTCAAGGAGA CCAGAGCGTTTGCTTTTGTAGCATGATTCCCGCTGGAGAGTCCCTGGAGATGTTG	55.0
rs820371	3	123404711	GTACACTACCCTGTAATGTGCTTTGAGTCTTATTTTTAGAAGTTTCTGCAGCACTTTATCTTTCCAT CTGGCCATGAGGAAAGGGGCTGAAGTCCCCCTTTCAAACTAGGCACACCCCAG	45.8
rs6437783	3	108172817	GCAGCTCTTGACCTACTATAAGGCAATGAGATTAGTTGCACTGGTTGAGGCACACTATTACCCTAG CAGTTGGGGAGCAGGGTGGGCATATAATACCTTTCTTCCAGAATCTTCTGGAAT	45.8
rs9809818	3	71480566	GTGTGGTTTTCTCAGCGACTTAACTTTTCACTCCAGTGAATACTCCTAATAAGAGCTGGCCTGGAT GACTCAGTGCTCATACCATGCTAGGCACTTAGCTAAGCTACGAGCTTCAGGCAC	47.5
rs12498138	3	121459589	AAATGTAAGAAATTAAGAGTTACATAGGATTTGCGAGAAACAGATAAATATTGAAGATAAGTAAT GTGAATAACAGGATTCCCTGAAGAAGAAAAGCAAAGCAAAGAAACAGCACAAATA	30.0

rs4683510	3	140285115	GCCCAGGGGTCTGCTGCCTGGCCCACATGTCCCTTTTGTAAACCCTGACCCAGTGTATGCCCATGT CTATCATACCTCACCTCTGATGTCTGTGACATGTCTGGGAAGGCCTTCTCCAGC	55.8
rs7623065	3	22385375	TACAGTCATTCTCACACTTGAATGTAAATCTAAATTACCTGGAATGCTTATTAATTCCCCAACTACT GGGCTGTACCCCCAGCATATTTACATTTTACCACATCCATTAAAAAAAATGAA	34.2
rs10012227	4	18637315	AAGTAAGAATCATGCTTGACTGGAATGAGGGAATCTTGCAGGAGATAAAAGCACAGATGGGGAT GTAAGATGCTGTGATAAGCACAGATCCTCTTTCTTTTGTGTGTTGGAGGGGAGGGA	44.2
rs1229984	4	100239319	TCTTTTCTGAATCTGAACAGCTTCTCTTTATTCTGTAGATGGTGGCTGTAGGAATCTGTCACACAGA TGACCACGTGGTTAGTGGCAACCTGGTGACCCCCCTTCCTGTGATTTTAGGCC	47.5
rs4540055	4	38803255	TCTCGCACTCATGCACGGGGTCCTGTGTGCCTCTGATCACTTTTGAATACCACGCCTGACTGTGCT GATCACTGCTTCCATGAACTCTGGAGTTGGCTAGGAGGTGGAAAGAGCCTTGAT	52.5
rs1509524	4	125455038	GGGGAAAGAAGATTGGAATAGAAAATTTTCTTCTAACTCTGTCAATATTACATTTAGGAGACCTCT TATCAGACTGAAATTTCAATCTCCAATAAGACCACAAGAAAAAGAAAATGTCTG	32.5
rs6875659	5	175158653	CCACCCACCACCATACACACACACAGGCACACACTCACACTCACCAGGCCCCAGTGAAAAGGAGG CCCCTACACATCTCACGTGACTAAATAGCACGTTTCTAGAGCCAGGGAAGCTTGC	55.0
rs16891982	5	33951693	CACATAGAAATATCAAATCCAAGTTGTGCTAGACCAGAAACTTTTAGAAGACATCCTTAGGAGAG AGAAAGACTTACAAGAATAAAGTGAGGAAAACACGGAGTTGATGCACCAAGCCCC	36.7
rs4704322	5	75822474	ACAATAACTAGAACTGGTGTTGCCATCCACCCCATTTGGGCATGAGCTTTTTTCTTCCACCGGAGA CGTAGGTCCCCAGATGATTGATTGGCTGCAAAGTATGGACATGTCTTGATTGTT	45.8
rs6886019	5	170245846	GGATTTTGCTTTTCAATAACAAATTATTGAACCAATTCACAAAAACATATCTTGAAGGATCGAAAA GCCCTCATTCATAAACCAAATCATATCTATGGAACAAAACAAAGGAAATGGGAC	31.7
rs10455681	6	69802502	AAGCAACATGTGTGAAAAAGATTTTGTGTGCAAGTTACTGGCTGGGCACTGAGGACATACATATT AATAAACACAGATCCTTCCCTTAAGAAGTTACAGTTAGTGATGAAAACAGGAACG	38.3
rs2080161	7	13331150	AATTTTGTTCCTGCTGTAACCATTTGCTTATTCTTTTAATTGTTCCACTATGAAAAAATTTGATGAG TGCAAAAAAAAAACAAAAAAAAAACCAAAAAAACAACAACCCATAAAGTCCAG	26.7
rs798949	7	120765954	GTACCCTGTAGTGCTCCAGGTCAGTATGCCAAAGGCCTCATGTGAGCTTCTACCTTGTGGTGAGAG ATCAGTGTGTGTGTTGTTGTTGTTGTTTTCATTCATAGTAACCAATCAATAA	43.3
rs1871534	8	145639681	GGAGGGCATGGCCTGGGAGTCCATGGTGGGGAACGGAGGGCCAGGGTCGCGGGTTTGTGGGGGC AGACCTTGGGCGTCAGATGCAGGACAGCGTCCCCAGTGAGTGCACCCACTGCCAG	68.3
rs2409722	8	11039816	GGTGACGTCGTAGGATGGGGAATTTGAAAGAGCACTCTTCTGACTTAAAATACCACACAGTTCCTT TATTGGAATAATTTCAGCCTTAATGAGACTGGGGTTGTCTTTTTAAACCCTGTC	40.8
rs7832008	8	98358246	ACTTGGTGGTTAGCCCACCCAGCTGTGCTGCTGGCCTTCATTGGTAGCCACTCCTGACACGAGGGT TTCAGTCACATCTTTCTTACACTCCCTATTTACACAACAGGGTTATTGAGATAA	48.3
rs2789823	9	136769888	TCCTCCCCGCAAGCTTCTCAATGCACCACTACTGTCCAAGCTGCTGAGTCAGTGCTGAGGGCCCCG TGAAGCTCACCTCCCCAAGAAGGCCCCAGAGTACCGGGGTGTGAACAGAAGCCC	60.8
rs10811102	9	1911291	AGCCTCTGAGATGGGATTCCTGTGTCTCTGACATTTATTGATTCCTCACTTTGTACTACTGTGTAAA GTAAGTTACCTGCATTATTTTATTCAATGCTCTTTAAGGCCTTTCCATATTCA	36.7
rs10970986	9	32453278	AGCAAATAAGGATGAGGGATGCCCTTCATGTGTATTTACACACCAGATTTTTATGGTACATAGGG AGTAATCTAGGGAAAAGATATATTAGGCATGAATCTTAGCTGTGGTAGAAGCACA	38.3

rs16913918	9	3074359	TCACTGCAGTGTGCACAGAGTCAGATCCAAAGAGCACTCTCCATGCTCTTATTATGAATGACAATA AAACCACACTCAGGGAGAGAAGGAAGAGGAAACAAATAAGCAGAACCTCAAATG	42.5
rs7084970	10	119750413	GACTGATAACACCTAAAGAGATCAAAAGAAGACTGAGGCAGAACACTGTTCACAATTAAAACGA ATAAAATTGAATTTCATAATCCTAAACTGCAAGTTCCTCCAGACGGGGAGCGAGTC	39.2
rs4749305	10	28391596	GTAAATAGTAACTCCATCTTCATAGTTACTAAATCCAAAAACCATACACCCATCCTTGGTTTGCCA CTTTGTAGGGAAATCTCCTCAGTTCTACCATCAAAATATATCCAGAATGGGACA	37.5
rs2274636	10	27443012	GTCGTCCCCTGATCCACTCCTCAATCCTGGGATTCGCATTGAGTACATCACTTCATTTCTAGAGTCC CTTCTGAGAAATCGCTTCCACGAGCGCAATTTGCAAACAGCCACGGGCAGGGC	52.5
rs174570	11	61597212	GATACAGAAATTGGGGAAAGTGAATAGGAGAGAGGCAGAAAGGAGGGATGAACTTGACGTAGAT CATTCCACCTGGAGGTCTAGGATGCCCAACTGTTGGTGGGCTTTTGCTCTCTAAGC	48.3
rs3751050	11	9091244	GCAATAGAGAGCTCCAGTGTATTGGGAAGGCTCCCAACTCGTTAGGAGAGTTGAGACATCATATC TCTTGGGTGACAGAATAAATTTTTCATGTCTATTAATTGGCCTAGGTTGACTTTA	40.8
rs5030240	11	32424389	TTAGCTGTCAAGACAGATTTCTAATTTTTACCCTTCGATGGCAGAACTTTAATGTTGTGTCCCTAGA AATCCTTCAGCCCTAGAAAATGTGGAAAGTAGGCCAGGGGCAGTGGCTCACT	43.7
rs1924381	13	72321856	GGCAGAACTCAGATTTCCTTGCTTCCTCCCATGGACCCAGTAGAGACACTCTCAAGGTAAATGTAG CTCCCAAATTTATCTAAAACAAACATGTAGTGAGGAAACTCACAGAATCTCCTA	42.5
rs9522149	13	111827167	CTGTACGCTGGGGGTCGTAAGCAGAAAGGAGAGGAAACACCGGGAGGTCCTTGCAGCTCCTAGG AAGCAGTTTGCTGATAACTCAGCTAGGCGTGCGAGGACTAGAAGTTGCTGACAGTC	55.0
rs721367	13	95546650	AAAAAGAATCCACACATAAAAATCTCTCAGACTTTCCTTCCCGGTATTTTATTTGGCAAAGGACCT ATGGGGTGAGAAAACTGTACTAAACATTACAGTCAGTCTATGGTTTACATGAAA	35.8
rs730570	14	101142890	CTGACGGAGACACTCCCCTACTCCAGGCCTGCAGCACTCACCTGCATCTCACACTGCGATGCAAA ATTTGTGTGATTAATGGGTTCAAATGTGCCCCCCTCCCCTGGACTTTACAGGGTG	54.2
rs7151991	14	32635572	GGTTATCCATTTTCATAAAAAAAGAAGTGGCTCAAGTTCAGCATATATACAGACCCATGGACAAT GACAAATGATCAAGGATTAAAATATTAGGAACACAGATGCCTGGGGCAGAGACAT	37.5
rs10483251	14	21671277	CTTGGCTATTAGTCTTGAGTAGCACATTTCTTCCTAGGGAAAAAGTTATGTGACCAGATTGTACCT TAATTGTGGGTGGGATATTCAGAAGCTGCTAAGTATGAATGTGCTCAAGAAGAG	40.0
rs12434466	14	97324289	TAGAAATGGAGTTAATACAGTGAATGAAATATGTAACAATTCCTGTCCTCTTGAAATTATTGTTTA CTTCTGTTAGTGTCCTTTTCTTTCACTTAGTAATGTATTTGTGAAATTCATCCC	30.0
rs1834640	15	48392165	GACCCATACTTGTAGTGACTGAGACACAGTGACATTATATCACAACCTCAGAAACCACAACATAA ACCAAGGAATAATCAATGCCATAGTTTTTAATAGTGCAACTAGAAATGGATTGTC	36.7
rs12594144	15	64161351	TCTCCTACAAGACCCTCTCCTACAAGACCACCCACCAGTCTGGGTCCTAATTGTTTGTGACTCTTTT TCAGGGTGGGAGCAGGGTGGTTATGGAATGACCATGGGTCCCGGCATGTAATA	50.8
rs1426654	15	48426484	GCTTTGAGTATCTATTGTGTTTAGTTGTAAAGACATACTCTTTCACTTTATTAGGCATAACAATCAT TTCATTTATGTTCAGCCCTTGGATTGTCTCAGGATGTTGCAGGCACAACTTTC	35.8
rs3784651	15	94925273	AAGGGCCTCCGGGGCCTGTCTTACACAGCCAAGTACCCAGGCCTGAAAGGCCAGAACTCATAAAG GGAATACTGACTTTGTAGCTTCTGTGAATGAAAGTGGGCTCTGACCTCTTCTTCA	50.8
rs6494411	15	63835861	TGCTTAGTTATATGGATTTATTCGTCCCATATAATATGTGTAGCCACATGGACACATGTTCTAAAA GGTTTTAAATTCTGCAAGATAGTGACGATGATGTAAACACAAGAAAAGGGCAAA	34.2

rs881929	16	31079371	CAGAGGGCTTGACCCAGTGGTTCTGAGCTCTGGCCCTTCTACTGCCTCTTGCCAGCTCTGGGTCTC AGCCTTCCTATCTGTGAGTTAGACACCAGGTAGCTGGAGGGGAAATCCCTCCTC	57.5
rs17822931	16	48258198	TCGCTAAACCTCTGAAGCCTAGAGTCCCCCAAACCTCACCAAGTCTGCCACTTACTGGCCCGAGTA CACTGGCAATGCAGAAGCAGATGCCCAGAAGTGCATCGAAAATCAACCTTGTTC	51.7
rs16946159	16	48459558	AGTCAATTATCTTTCTAATACAGAAAATATTTAGAAAGCCACATATTTACACTGAATATGGGCTAA AACAGCATTTGTATTTTCTTTGACAAAGCTCATATTTAATGAAGTATGTTTATG	26.7
rs4792928	17	42105174	TGGATAAGATGAAAACACAGAAGACAGTTTATCAAAATTAAGAATAACACAACATTGGGATGGA CAACTAGTAGACTGCGATACAGAATCAAGATTTTTTTAAAGCTCTTAATAGGTTAA	30.8
rs8072587	17	19211073	GGGTGATGCCTCTGCTTGGCAACCTCACATGGTAGAGAAGTTCCCAGTCTCCTGCCCGGCCCCACC TCAGCCCTCTGCCCAGAAGCAGCCTCTTTCTACCTCCCCTGGTTCTTCTGTGGA	60.0
rs9908046	17	53563782	AATCTTAGACAAAATTTCAGAACAGATGAGGTTTACCTTGGCATGTTCCTCTCTCTATCTCTGGGT GCTCCTACCTCCATCTGCATCTCTCCTCATGACCCCAAATGTCTAGTAACTTTA	42.5
rs1369290	18	67691520	GATCACTCTCACCAGAACTCACAATGCATGGCCCACCACAGGCTCTTGATAAAGTGTCTATCAAGT GCATGAGCGAAAGAGGAGCCCTAGTGATATCAAGTATCACCAGACCAACTGTTA	46.7
rs310644	20	62159504	TATGAATGTTGCAGAAACTCTATAAAATTCCTAGAAATCTGATACGTTATCCTATGATATATATCC TATCTCACATGCTATTCCTAGAAATCTGGTGCAATGTTATCAGTCATAATTTTG	30.8
rs2069945	20	33761837	CTGGACTTGACTGCCTACGTGCAAACCTTGGCTCTGCTACACTATCTCTGTCTCAGTTTCCCATGTA GACTGGGGTTAATAATAGTAGCTATTGCATTAAGCCACTGGGGAAAGGCACAA	46.7
rs6054465	20	6673018	CCAAGGGCCATAATCTTTTATGGCCTCAGGTTCTCCACATGCAAAATCAGGATAATAATGTTCTCT ACCTCATGAGGTTATTTAAAGGAAACGGTGAGATCATGTGTTTTGATAGAATAT	37.5
rs715605	22	30640308	TGCAGGGATGGAAGTACTTGTGTGTGGTGCCCCAGCTAGGGCTAGACACCGAGTTTTCCCTTCTGT CCCCTTAGGGTGGTGATGATGATGATGATGATAATGATGACTGCGTGCATGGCT	51.7
rs1557553	22	44760984	AGTCTCAAATTTAATACAAGAGCCGCCTGGACTTTGCATTACCCAAAGCCCCTGGAAAAACGACG TGAAATAGCTGCAGTTCCTTGCAAGGGGGGAGAGTGAGTCTGCCAGACAGGAGTG	50.0
rs8137373	22	41729216	GGCAGCAGCAGGCCTCCCTCCTCCAGAGCTTTGCAGCACTTTCTTTCATTCATTCCATAAGGAGGC CCACAAAACACTCTCGGCCCTGGGCCTGAGAGAGCTGCGTCCTTGCCCTCAGGG	59.2
rs4892491	X	73422412	ATGGATTACCCTACAAAACATCCTAAGTGAAAGAAACCCATGATAAAAGGCATCTATTGTGTTAT TCCATTTACGTGAATTATCCGGAACAGGGGAATCCATAGAGACAGAACATAAATC	36.7
rs11156577	X	153660041	GAGAGTGCTGGAGGGAAGAGGTGAGTCTCCCACCCCACCCCCACCAACACACAGTTGCGCTCCAC CATACTGAACCTGACTCTCAGTGAGACTTTGCTGGCCCTGAGAATGCACGGGGAA	59.2

Supplementary Table S2. Bait sequence (120 bp) and GC content for the 23 phenotype SNPs included in the hybridisation enrichment panel. Position of the targeted SNPs are shown in red. rs16891982 is also included in the ancestry SNP set. SNPs are listed in the same order as in Table 2. Three SNPs (rs1805005, rs1805006 and rs2228479) are on a 97 bp segment and were targeted using a single bait (highlighted in light grey). Six SNPs (rs11547464, rs1805007, rs201326893, rs1110400, rs1805008 and rs885479) are on a 64 bp segment and were targeted using a single bait (highlighted in dark grey).

rs Number	Chr.	Position (GRCh37/hg19)	Bait Sequence	GC content
rs16891982	5	33951693	CACATAGAAATATCAAATCCAAGTTGTGCTAGACCAGAAACTTTTAGAAGACATCCTTAGGAGA GAGAAAGACTTACAAGAATAAAGTGAGGAAAACACGGAGTTGATGCACAAGCCCC	36.7
rs28777	5	33994716	GGAGTTCCATGACTTTCAAAAGGCTTCCACTCAGTTGATTTCATGTGATCCTCACAGCAGCCTCT GAGTGGATGGGACGACCCCTGTCCTATGCGACTCTTGGAGGGGGAGTCCCAGTCT	53.3
rs4959270	6	457748	TGGGGTTTACGATTCAACATGAGATCTGGGTGAGGAACACATCCAAACTATGACACTATGCCAC TTCCACAGGGGGTAAGAACACTGACATGGTGGTTCTTTTCCATCCTTTGTGCTGTT	45.8
rs12203592	6	396321	TCAGGCTTTCTTGATGTGAATGACAGCTTTGTTTCATCCACTTTGGTGGGTAAAAGAAGGCAAAT TCCCCTGTGGTACTTTTGGTGCCAGGTTTAGCCATATGACGAAGCTTTACATAAA	41.7
rs683	9	12709305	AGTTATTAACTGTATTTTCTTTCACTTTATTACCTTCTTTCTAATACAAGCATATGTTAGCATTAA AGTTCTAGGCATACTTTTCAAAGCTGGGAAGACCCTTTCAGAATCTTTTCAATG	30.8
rs1042602	11	88511696	TTTATGACCTCTTTGTCTGGATGCATTATTATGTGTCAATGGATGCACTGCTTGGGGGATCTGAA ATCTGGAGAGACATTGATTTTGCCCATGAAGCACCAGCTTTTCTGCCTTGGCATA	43.3
rs1393350	11	89277878	ATCCCCCTGATGCGTGCATATCCACCAACTCCTACTCTTCCTCAGTCCCTTCTCTGCAACGAAATC TGTGTGGTCTTTTACAAATGATAATCTGTTCGTGTTATCATTCACCTTCCCAGA	45.0
rs12821256	12	89328335	AGTTGTGTGGCAGAAGTTGAAAATTAATTAAGCTCTGTGTTTAGGGTTTTTTTCCTTTAGTGTGCC GTAGTAACATGCCCTTGGCTCCAGGGAACTTTATGACTAAGTTGGAACAAAGCA	40.0
rs2402130	14	92801203	CAGTGCTGCTGTGTGACACCACATGCTGTATGGAAGTATTTGAACCATACGGAGCCCGTGGTAG CTGCCATCATCATCATCATCATCATCGTCATCATCATCATCATCATCGTCATCATC	46.7
rs12896399	14	92773663	AGTATCCTATATTTTATCTGGCGATCCAATTCTTTGTTCTTTAGGTCAGTATATTTTGGGGTCTCTT TGTCACAGCAGATTAACCTTCCTCATCAATACAACATCAAGACCCAGGGCTAA	38.3
rs12913832	15	28365618	GTTCTTCATGGCTCTCTGTGTCTGATCCAAGAGGCGAGGCCAGTTTCATTTGAGCATTAAATGTC AAGTTCTGCACGCTATCATCATCAGGGGCCGAGGCTTCTCTTTGTTTTTAATTAA	44.17
rs1800407	15	28230318	CCTGCTCACTCTGGCTTGTACTCTCTCTGTGTGTGTGTGGCCAGGCATACCGGCTCTCCCGGGGA CGGGTGTGGGCCATGATCATCATGCTCTGTCTCATCGCGGCCGTCCTCTCTGCCT	61.7
rs1805005	16	89985844	CAAGAACCGGAACCTGCACTCACCCATGTACTGCTTCATCTGCTGCCTGGCCTTGTCGGACCTGC TGGTGAGCGGGAGCAACGTGCTGGAGACGGCCGTCATCCTCCTGCTGGAGGCCGG	63.3
rs1805006	16	89985918		
rs2228479	16	89985940		
rs11547464	16	89986091	AGCCTCTGCTTCCTGGGCGCCATCGCCGTGGACCGCTACATCTCCATCTTCTACGCACTGCGCTA CCACAGCATCGTGACCCTGCCGCGGGCGCGGCGAGCCGTTGCGGCCATCTGGGTG	67.5
rs1805007	16	89986117		
rs201326893	16	89986122		

rs1110400	16	89986130		
rs1805008	16	89986144		
rs885479	16	89986154		
rs1805009	16	89986546	GGCTGCATCTTCAAGAACTTCAACCTCTTTCTCGCCCTCATCATCTGCAATGCCATCATCGACCCC CTCATCTACGCCTTCCACAGCCAGGAGCTCCGCAGGACGCTCAAGGAGGTGCTG	56.7
rs2378249	20	33218090	GACCTCAGTTCTGGAGAAAGCTAACTAAGGGCACAAGTCTAGGAACTACTTTGCACAGTAGTGG GCTGAGGAGAGGTGTGGGCTGAAAAGCAATGAGATTAAATAAAGTCTCCAGACTGA	45.8

Supplementary Table S3. Bait sequence (120 bp) and GC content for the 35 Y-chromosome SNPs included in the hybridisation enrichment panel. Position of the targeted SNPs are shown in red. SNPs are listed in the same order as in Table 3.

rs Number (mutation name)	Position (GRCh37/hg19)	Bait Sequence	GC content
rs2032595(M168)	14813991	TTTTGCAGAGAGCTTGGAGATAATTCTGGTGGCTGTGTGGAGTATGTGTTGGAGGTGAGTTGCT AGCTGAAGAATTAAAACAATAGTTTTAGCAGTTTGGGTAAGAGATGTTTACAGAAA	39.2
rs3848982(M145)	21717208	TCTAATTTTTATAGCGGCATACTTGCCTCCACGACTTTCCTAGACACCAGAAAGAAAGGCGAG AGCCAGCCTTAGCCTAATCAAGAACCATGATCCAAAAAGGAGTGGGAGGAACCTAGC	46.7
rs2032602(M174)	14954280	TACTCATAATGTCCTTTTTAATGTATCAAATCGCTTCTCTGAATACCTTCTGGAGTGCCCTAGTG CAGAAGTGAGGGGTGCATTTGCAAAACTTATAGTGTTTATTGCACACTTTTCCTT	38.3
rs371443469(V36)	6814246	CAAATCACCATTTAGGAGACTAGGCAATTTTTCCATTGAAATTTGGATAATTTCTTTTTGAAAA GCAAAATTTTGCAATATAAGTTTCAGCAAAAACATCTCCTTATTAAACAATTATAG	26.7
rs9306841(M96)	21778998	CCTATTATGAGAGACCTGTTTTCCAAGTTACACATCACAGCTCACTTTAAAGTGCTCTGTGAGA GGGCAACAGGCCAGAAGGAACAGTTCTCCTAGCTTCAAATACTTGCCTTTACACAG	44.2
rs9786025(P170)	15021522	GTTTTTGGAACAGTTTGTATAATGTTGAATGATGGATAGTTGTTTCCTTTGGCAAACTGAGCCC AAAGAATCTGAGTTTTGTATAGTGGTTTATCTGTTTGTAAACATTTATGTTTTGAA	31.7
rs2032666(M216)	15437564	TTATATCCTCAACCAGTTTTTATGAAGCTAGAAAAAAATTCCTTTATTAAAGAAATGTAACATT CAACAGGTATACATAACTAGCAGTGTCAGAATTCAGATTTAGAACCATGTTTACTA	28.3
rs35284970(M130)	2734854	TTTAAAAAACAAAAAAACAGAAACACTGTCAATCTGATTTGATCTACCTGCCCTTTCCTGTCTT GTTGCAGCCCATGTATCCCTGGAAATCAATTTGTTCCCTTTAATTTTACATTGATA	45.0
rs2032668(M217)	15437333	TCATATTTTAGGTTAAGATTCTTTAACTTGTGAAGGAGAATGAAAAAGTTGGGTGACACAAAC TCTTCAGAAGGAAAAATACATAAAAATTATTTTGATGAAAGCCACAGCAGCTTTATC	30.8
rs868363758(M347)	2877479	CACAGCTGGTAGTCCCACAGTCTGTGTAAGTCTGGCTGTCTGAGGTTTTTATGGACTTCAGAAG GCAGGAAGTGTGTGCTGACTCGTCCAAGGGCAGCCATGTGCAGGCCTGGAAAAAGC	53.3
rs9786706(U13)	14698928	GGAATGGATAGTAGTGGTAACAGCTCCTGGTGAGCAGAGGTCAGAGGTGCTGCTAAGCCTTCT ACAATGTACAGGACAGTTAACCAAAGCAGCAGAACAAAATGTATCCAGCCCCAAATA	46.7
rs2032636(M201)	15027529	GTTACTACTTGAGTTACTATATTAGTGCAATTAATTACACAACTATATATAGTAATTAGTTTCT CAGATCTAATAATCCAGTATCAACTGAGGTTTTTCGTAATAGGTACTTAGTGTTGG	29.2
rs13447371(M282)	21764431	TGCATGTGATCAACTTCTTTCCCTCAACATAGTTATCTCAGAGTTTTGTTGAAGAAAACTTTAT CATATTGAGATTTTGCTTTCCTATAAAGCAAAGTTGAGGTTGCACAAACTCAGTAT	33.3
rs2032673(M69)	21894058	GTACTCCTGGGTAGCCTGTTCAAATCCAAAAGCTTCAGGAGGCTGTTTACACTCCTGAAATAA AATATATTTCAGCAAGACAAAGGGAATAAAGATCCAAAAAAACAGGAGAGCTAAGGG	40.0
rs17250163(P126)	21225770	CTCCAGACAAATCTCGTCTCATAAAATGGGAAGCTGTCATACCCTCGGATCTCCTCCAGCCCTA ATACTCAGGTGACTACTTGGTAAAGGTGCTCTTCATTGAGCAGGAGAAACCATCAG	47.5
rs9341301(M258)	15023364	ATGGGACATAACTAAGAAACTAACAATGGGCTTGCACTAGCTACAAGTTCAGCTTAAAAATTG GGAACTTGGAATCCCTCTTAGTCATAGCTTAAAAAAAGACTCATCTTAAATAATTTA	33.3
rs13447352(M304)	22749853	GTTTGTAACAAACAGTATGTGGGATTTTTTTAGATGTGTTCAATTTGAAAGTAACTTGTGAAAC AACTGGTGATATTTTGGTATAAGACGTTTTGAAAGTTATTTGTTTATTTCTAAGGA	27.5

rs9341313(M267)	22741818	AATAATGATTCCTTGGATATACCAAGTCTGGATAGCGGATTCGATGGAAGCATTTTTGTAAAT ATACGTTCAGTATTTTGTGTGGAAGAACACAATCTAGCTGATGCCTGCAATCCCAG	38.7
rs3900(M9)	21730257	ATAAAACTTTCAGGACCCTGAAATACAGAACTGCAAAGAAACGGCCTAAGATGGTTGAATGCT CTTTATTTTTCTTTAATTTAGACATGTTCAAACGTTCAATGTCTTACATACTTAGTT	32.5
rs3902(M11)	21730647	TGTCTCTCCCTCCCTCTCTCCTTGTATTCTAACAGAAAGGTTTAGAACTTGCATAATTGGGAAA GAAGCTGTTGCCTGAACTTACTGGGGGATTCAGCATTGTCATTTTGGACATGTCAC	43.3
rs9341308(M272)	22738775	GAGGTACTTGGTGACAGTACAGTGCAGTCTTTCCTGGGCATTACTCTTTGCTCTCCCGAAAACC CACTAACGGGTTGTGTGTATAATAAGGTTTTATTTTATTTTATTTTATTTTTTACT	36.7
rs2033003(M526)	23550924	ATTTTTATTATTTATTTACTTAGAGGCAGGGTGTTGCTCTGTCATCAGGCTGAATCATACCTCAC TGTAGTGCCAGACTTTTGGGCTCAGAGGATCCTCCAACAGCAGCCTCCCAAAGTA	44.2
n/a (P308)	15409573	TCTTTTGAAAAAAATACTGAATGGCTACCAATACCCCCAAAGAGAAACTGCAGCAGTTTCAGA AGGCACTTACTTAATCATTTCAAAAAGCTTTGGATCTGAGACTTTGATATTTCGTGC	36.7
n/a (P256)	8685230	AGTTCTTGGTTTTCCCATTGACCTCTCTGAGGCCTTTCTGCCCTACACTAGATAGAAAGGGGTT GAGTAGAAGGCACAGACAAGTTGGGAGATGATTTCCTTTTCCTCAGAACTTCTGCT	45.8
rs2032631(M45)	21867787	CTTTTAACAGTAACTCTAGGAGAGAGGATATCAAAAATTGGCAGTGAAAAATTATAGATAAGC AAAAAGCTCCTTCTGAGGTCCAGGCCAGGAGATAGTAGGATTTAAGAAACAAACAAA	35.8
rs8179021(M242)	15018582	GTATCTTGAACTTATATATGTAAGCTTTCTACGGCATAGAAAGTTTGTGCAAAAAGGTGACCA AGGTGCTCTTGGCATTGGTCTTAACGTGTTTTTTGAAAAAAATCTATTTTAACGTA	33.6
rs2032658(M207)	15581983	CCCTGAAGAAGGAAAAAACGTTACAACTATGGGGCAAATGTAAGTCAAGCAAGAAATTTAGA AAGAGAATAACAATACCTTTTGAATAATCTTCCAACAAGAGGTTGAAGTGACCTAATT	34.2
rs17250535(M420)	23473201	CTTTTGGGATTATTTATTGTTATCTAAACAAGATTTTTTTTTTTCCATTTACAGCAAAATGGTGG AAGCAGATTGGCCTGGCAAACTTTTCATTGCTGGCCTCCATTTAGAAACCAATGA	34.2
rs9786184(M343)	2887824	GAAAACTGGCCACCCTAGCCTTTTAAATATGCAAATGCAGAGTGCCCTCGTGTTCCAAACACC TGGAGATATGTGGGGGTGGCTATGCTGCCAGGCACGTGTTGGGGAAAGAGCAAGAGG	52.5
rs9786153(M269)	22739367	TAAAGTGGATTCTGTTACATGGTATCACAATAGAAGGGGAATGATCAGGGTTTGGTTAATCCT GGTAAATTGAAAACAATTTTTTTTTTATCATATGTGCCTCAGAAGGCACACAAAAGA	34.2
rs9786140(M412)	8502236	GGATAGAGAGTGTGAGAAGAAATAAGGTGAGATATGGACGGGGGTACAATCTGATGAGGCAT GGATAGGGTCCACTTCACCTGTAAAATACATGAAGATAGATGACATAGAAGGTGCTTC	43.3
rs9341278(M231)	15469724	ATTCATATTAATTTAGTTAATCATCATTCAATTAATACCTAAAAAACAACATTTACTGTTTCTA CTGCTTTCGAATTGGGGGAAAGATCGTCAAAGAATTCATACCTGTAATTTCTGTGG	29.2
rs13447361(M324)	2821786	TTTAAAAAACAAAAAAACAGAAACACTGTCAATCTGATTTGATCTACCTGCCCTTTCCTGTCTT GTTGCAGCCCATGTATCCCTGGAAATCAATTTGTTCCCTTTAATTTTACATTGATA	34.2
rs11575897(M176)	2655180	GAAGATGCTGCCGAAGAATTGCAGTTTGCTTCCCGCAGATCCCGCTTCGGTACTCTGCAGCGA AGTGCAACTGGACAACAGGTTGTACAGGGATGACTGTACGAAAGCCACACACTCAAG	52.5
rs13447354(M307)	22750951	GTGCTCAAATCCTTTTTGAAGGCTACATGGAAATGGTTGGCTATTGAGTTAGCATAATCAGTCT GCCTACCATACTTAAAGTACCTTTTGTATATGTGCTAAGTGAGAATTAAAAATACC	35.8

Supplementary File S4. Details of the 368 reference samples from five population groups across the 1000 Genomes and HGDP-CEPH datasets used for comparison. All genotypes are presented in forward orientation (provided as an electronic file on USB Drive).

Supplementary File S5. SNP details and population-specific/pairwise Divergence values (I_n) of the 67 ancestry-informative SNPs in the custom enrichment panel (including four tri-allelic SNPs). SNPs are ranked according to their population I_n (highlighted in grey) inside each population group.

SNP Details		Population-specific Divergence (In)					Pairwise Divergence (In)									
Group	SNP ID	AFR	AMR	EAS	EUR	OCE	AFR/AMR	AFR/EAS	AFR/EUR	AFR/OCE	AMR/EAS	AMR/EUR	AMR/OCE	EAS/EUR	EAS/OCE	EUR/OCE
AFR	rs2814778	0.670	0.112	0.130	0.129	0.135	0.654	0.660	0.660	0.633	0.000	0.000	0.001	0.000	0.000	0.000
	rs1871534	0.637	0.109	0.126	0.125	0.082	0.621	0.626	0.626	0.599	0.000	0.000	0.000	0.000	0.000	0.000
	rs6875659	0.598	0.099	0.130	0.072	0.093	0.609	0.634	0.546	0.621	0.001	0.005	0.000	0.010	0.000	0.007
	rs2789823	0.593	0.107	0.123	0.121	0.080	0.577	0.582	0.582	0.556	0.000	0.000	0.000	0.000	0.000	0.002
	rs1369290	0.455	0.230	0.162	0.103	0.100	0.277	0.484	0.413	0.459	0.379	0.322	0.376	0.009	0.000	0.003
	rs310644	0.369	0.107	0.170	0.081	0.194	0.477	0.540	0.419	0.003	0.006	0.004	0.426	0.019	0.487	0.370
AMR	rs10483251	0.074	0.444	0.074	0.002	0.000	0.529	0.000	0.032	0.051	0.536	0.348	0.301	0.035	0.055	0.002
	rs12498138	0.098	0.431	0.022	0.024	0.016	0.518	0.024	0.022	0.024	0.386	0.392	0.387	0.000	0.000	0.000
	rs2080161	0.143	0.431	0.002	0.019	0.027	0.621	0.103	0.045	0.027	0.292	0.408	0.462	0.014	0.040	0.016
	rs174570	0.172	0.344	0.000	0.026	0.023	0.582	0.115	0.053	0.212	0.247	0.363	0.137	0.015	0.020	0.069
	rs1557553	0.064	0.320	0.000	0.050	0.004	0.403	0.039	0.000	0.020	0.219	0.385	0.270	0.033	0.003	0.015
	rs12402499	0.068	0.317	0.064	0.006	0.037	0.360	0.000	0.026	0.000	0.358	0.236	0.334	0.026	0.000	0.016
	rs7151991	0.021	0.309	0.020	0.008	0.089	0.292	0.000	0.002	0.033	0.292	0.257	0.455	0.002	0.033	0.047
	rs10012227	0.092	0.307	0.011	0.041	0.012	0.445	0.096	0.006	0.019	0.159	0.368	0.310	0.056	0.032	0.004
	rs8137373	0.171	0.232	0.202	0.038	0.200	0.015	0.407	0.206	0.475	0.521	0.301	0.593	0.048	0.007	0.084
	rs4792928	0.161	0.194	0.093	0.151	0.009	0.412	0.276	0.000	0.063	0.020	0.411	0.198	0.274	0.000	0.062
	rs647325	0.008	0.149	0.001	0.111	0.024	0.073	0.002	0.103	0.044	0.100	0.318	0.215	0.075	0.026	0.013
EAS	rs1229984	0.089	0.072	0.356	0.074	0.018	0	0.368	0	0.018	0.362	0.000	0.015	0.354	0.269	0.013
	rs12434466	0.073	0.229	0.338	0.151	0.229	0.289	0.296	0.016	0.289	0.256	0.351	0.000	0.401	0.256	0.351
	rs4892491	0.004	0.053	0.332	0.061	0.019	0.022	0.255	0.022	0.006	0.394	0.000	0.005	0.395	0.328	0.005
	rs17822931	0.162	0.002	0.328	0.056	0.029	0.128	0.489	0.026	0.036	0.161	0.049	0.036	0.353	0.323	0.001
	rs12594144	0.256	0.120	0.264	0.086	0.001	0.429	0.560	0.043	0.148	0.016	0.254	0.101	0.370	0.186	0.041
	rs881929	0.172	0.009	0.261	0.000	0.008	0.069	0.461	0.101	0.067	0.216	0.004	0.000	0.168	0.220	0.004
	rs6437783	0.108	0.105	0.256	0.109	0.001	0.253	0.392	0.000	0.074	0.030	0.263	0.062	0.403	0.156	0.080
	rs6494411	0.083	0.083	0.251	0.105	0.007	0.258	0.417	0.001	0.046	0.029	0.231	0.098	0.386	0.216	0.034
	rs4683510	0.040	0.032	0.208	0.116	0.023	0.086	0.242	0.016	0.000	0.047	0.168	0.083	0.358	0.237	0.018
	rs4704322	0.124	0.004	0.202	0.031	0.002	0.102	0.349	0.020	0.096	0.090	0.034	0.000	0.222	0.096	0.030
	rs3827760	0.195	0.235	0.192	0.183	0.086	0.499	0.416	0.000	0.012	0.007	0.498	0.414	0.414	0.335	0.012
	rs721367	0.066	0.025	0.166	0.115	0.040	0.104	0.231	0.010	0.000	0.030	0.159	0.096	0.300	0.220	0.013
	rs4657449	0.188	0.136	0.133	0.138	0.213	0.377	0.349	0.002	0.513	0.001	0.335	0.022	0.308	0.031	0.467
EUR	rs1426654	0.094	0.066	0.119	0.631	0.077	0.001	0.003	0.616	0.000	0.007	0.593	0.002	0.658	0.000	0.631
	rs16891982	0.119	0.084	0.111	0.622	0.070	0.002	0.000	0.623	0.000	0.001	0.598	0.000	0.621	0.002	0.592
	rs1834640	0.140	0.005	0.122	0.490	0.090	0.125	0.000	0.616	0.000	0.115	0.255	0.113	0.601	0.000	0.598
	rs12142199	0.060	0.076	0.088	0.391	0.055	0.004	0.005	0.357	0.001	0.000	0.398	0.000	0.403	0.000	0.378
	rs9522149	0.086	0.051	0.081	0.370	0.050	0.003	0.000	0.370	0.000	0.003	0.335	0.000	0.369	0.000	0.345
	rs8072587	0.056	0.003	0.110	0.310	0.070	0.020	0.012	0.317	0.005	0.057	0.198	0.043	0.398	0.000	0.373
	rs820371	0.065	0.000	0.082	0.302	0.086	0.036	0.002	0.337	0.009	0.051	0.175	0.073	0.370	0.003	0.412
	rs7084970	0.092	0.046	0.026	0.300	0.084	0.003	0.012	0.402	0.235	0.003	0.351	0.193	0.296	0.150	0.034
	rs1924381	0.124	0.000	0.020	0.256	0.001	0.071	0.031	0.393	0.037	0.009	0.160	0.006	0.235	0.000	0.222
	rs4749305	0.009	0.142	0.120	0.246	0.089	0.156	0.121	0.097	0.119	0.004	0.434	0.005	0.387	0.098	0.384
	rs730570	0.017	0.005	0.098	0.218	0.128	0.024	0.025	0.201	0.063	0.094	0.092	0.153	0.338	0.011	0.425
	rs16913918	0.209	0.000	0.016	0.138	0.091	0.113	0.171	0.350	0.327	0.007	0.090	0.074	0.052	0.039	0.000
OCE	rs9908046	0.025	0.041	0.001	0.011	0.546	0.004	0.011	0.002	0.580	0.025	0.011	0.625	0.003	0.481	0.538
	rs2139931	0.030	0.008	0.023	0.003	0.470	0.004	0.000	0.031	0.540	0.002	0.014	0.481	0.025	0.524	0.363
	rs715605	0.002	0.034	0.041	0.005	0.458	0.018	0.020	0.001	0.402	0.000	0.013	0.517	0.014	0.522	0.425
	rs3751050	0.007	0.017	0.019	0.004	0.430	0.002	0.002	0.000	0.415	0.000	0.004	0.459	0.004	0.458	0.404
	rs6054465	0.052	0.003	0.001	0.006	0.421	0.018	0.042	0.016	0.540	0.005	0.000	0.407	0.007	0.336	0.418
	rs10970986	0.120	0.002	0.003	0.003	0.397	0.101	0.063	0.104	0.632	0.006	0.000	0.310	0.007	0.384	0.306
	rs16830500	0.058	0.018	0.081	0.090	0.394	0.006	0.149	0.005	0.534	0.100	0.021	0.455	0.192	0.164	0.595
	rs6886019	0.009	0.000	0.047	0.005	0.356	0.005	0.014	0.000	0.353	0.033	0.003	0.291	0.018	0.443	0.337
	rs2274636	0.079	0.009	0.010	0.007	0.346	0.025	0.088	0.030	0.502	0.024	0.000	0.367	0.019	0.226	0.352
	rs16946159	0.001	0.035	0.008	0.043	0.332	0.033	0.008	0.036	0.253	0.011	0.000	0.405	0.012	0.327	0.410
	rs1509524	0.001	0.005	0.005	0.008	0.314	0.001	0.001	0.002	0.289	0.000	0.000	0.324	0.000	0.318	0.332
	rs11156577	0.114	0.061	0.138	0.023	0.264	0.188	0.266	0.123	0.065	0.014	0.009	0.425	0.042	0.522	0.335
	rs7623065	0.163	0.157	0.119	0.001	0.262	0.371	0.302	0.106	0.045	0.007	0.102	0.579	0.062	0.502	0.256
	rs10455681	0.243	0.014	0.198	0.074	0.255	0.233	0.477	0.045	0.632	0.064	0.095	0.153	0.291	0.025	0.430
	rs10811102	0.232	0.045	0.066	0.029	0.254	0.292	0.310	0.075	0.620	0.000	0.092	0.103	0.104	0.091	0.344
	rs7832008	0.014	0.067	0.021	0.075	0.245	0.091	0.001	0.089	0.153	0.105	0.000	0.423	0.104	0.136	0.420
	rs3784651	0.011	0.071	0.069	0.139	0.232	0.089	0.016	0.136	0.150	0.173	0.006	0.415	0.233	0.076	0.494
	rs9809818	0.272	0.100	0.240	0.165	0.223	0.415	0.556	0.014	0.620	0.019	0.320	0.042	0.453	0.005	0.515
	rs10183022	0.305	0.085	0.003	0.007	0.209	0.419	0.231	0.248	0.632	0.040	0.033	0.045	0.000	0.155	0.141
	rs798949	0.087	0.131	0.102	0.007	0.175	0.259	0.209	0.024	0.035	0.004	0.137	0.437	0.099	0.376	0.110
	rs2409722	0.328	0.149	0.089	0.010	0.162	0.522	0.414	0.141	0.572	0.012	0.163	0.004	0.094	0.027	0.199
Tri-allelic	rs2184030	0.074	0.050	0.039	0.055	0.035	0.131	0.099	0.129	0.039	0.013	0.000	0.250	0.012	0.187	0.247
	rs5030240	0.074	0.177	0.172	0.154	0.123	0.129	0.237	0.213	0.066	0.356	0.356	0.085	0.002	0.457	0.432
	rs4540055	0.338	0.090	0.131	0.203	0.063	0.283	0.446	0.555	0.146	0.185	0.163	0.101	0.025	0.116	0.184
	rs2069945	0.106	0.067	0.095	0.023	0.047	0.193	0.214	0.074	0.117	0.001	0.039	0.383	0.050	0.403	0.269
Cumulative Values		9.678	7.265	7.334	8.056	9.705	14.472	13.971	10.879	14.76	6.707	11.718	14.139	11.648	11.13	15.028

Supplementary Table S6. Mean depth of coverage and standard deviation for 87 biogeographic ancestry and phenotype SNPs across twelve individuals. SNPs are ordered by increasing mean depth of coverage.

SNP	Mean Depth of Coverage	Standard Deviation
rs9809818	420	196
rs683	436	197
rs12498138	443	205
rs1426654	475	211
rs16891982	479	217
rs1229984	480	215
rs12821256	489	243
rs310644	490	220
rs5030240	491	231
rs6494411	492	217
rs4749305	511	238
rs12434466	513	227
rs16946159	513	237
rs6437783	529	254
rs1509524	534	241
rs8072587	535	246
rs1805005	543	286
rs4683510	552	275
rs10811102	552	244
rs17822931	554	258
rs1871534	560	282
rs16830500	560	253
rs881929	561	274
rs2139931	566	255
rs6054465	566	250
rs798949	566	249
rs9908046	571	260
rs4959270	572	267
rs6886019	572	262

SNP	Mean Depth of Coverage	Standard Deviation
rs7832008	574	262
rs1042602	575	263
rs3751050	576	280
rs4792928	576	302
rs12402499	580	257
rs4657449	581	284
rs7084970	584	272
rs2409722	585	283
rs7151991	586	270
rs7623065	586	271
rs2080161	587	244
rs174570	589	298
rs12896399	589	274
rs721367	590	269
rs4540055	594	298
rs4704322	600	290
rs10455681	601	265
rs28777	602	274
rs16913918	603	284
rs10012227	603	279
rs12594144	607	271
rs2378249	609	309
rs10483251	616	286
rs1834640	617	288
rs12203592	619	287
rs12913832	625	307
rs10970986	626	276
rs1393350	629	289
rs10183022	632	299

SNP	Mean Depth of Coverage	Standard Deviation
rs1924381	633	309
rs12142199	636	320
rs1369290	638	301
rs3784651	641	302
rs2274636	645	308
rs647325	652	316
rs2402130	657	295
rs2184030	658	319
rs2814778	659	333
rs8137373	660	328
rs3827760	661	327
rs9522149	668	327
rs2069945	670	329
rs2789823	672	340
rs820371	676	333
rs885479	677	354
rs1805009	681	312
rs730570	683	334
rs6875659	684	349
rs715605	695	334
rs1557553	713	351
rs1805008	724	379
rs1800407	738	363
rs2228479	758	389
rs1110400	760	406
rs201326893	765	411
rs1805007	768	413
rs11547464	777	423
rs1805006	779	398

Supplementary Table S7. Mean depth of coverage and standard deviation for sex chromosome SNPs (two X chromosome SNPs for all 12 samples [rs1115657 and rs4892491], and 35 Y-chr SNPs for six male samples). SNPs are ordered by increasing mean depth of coverage. M = male; F = female.

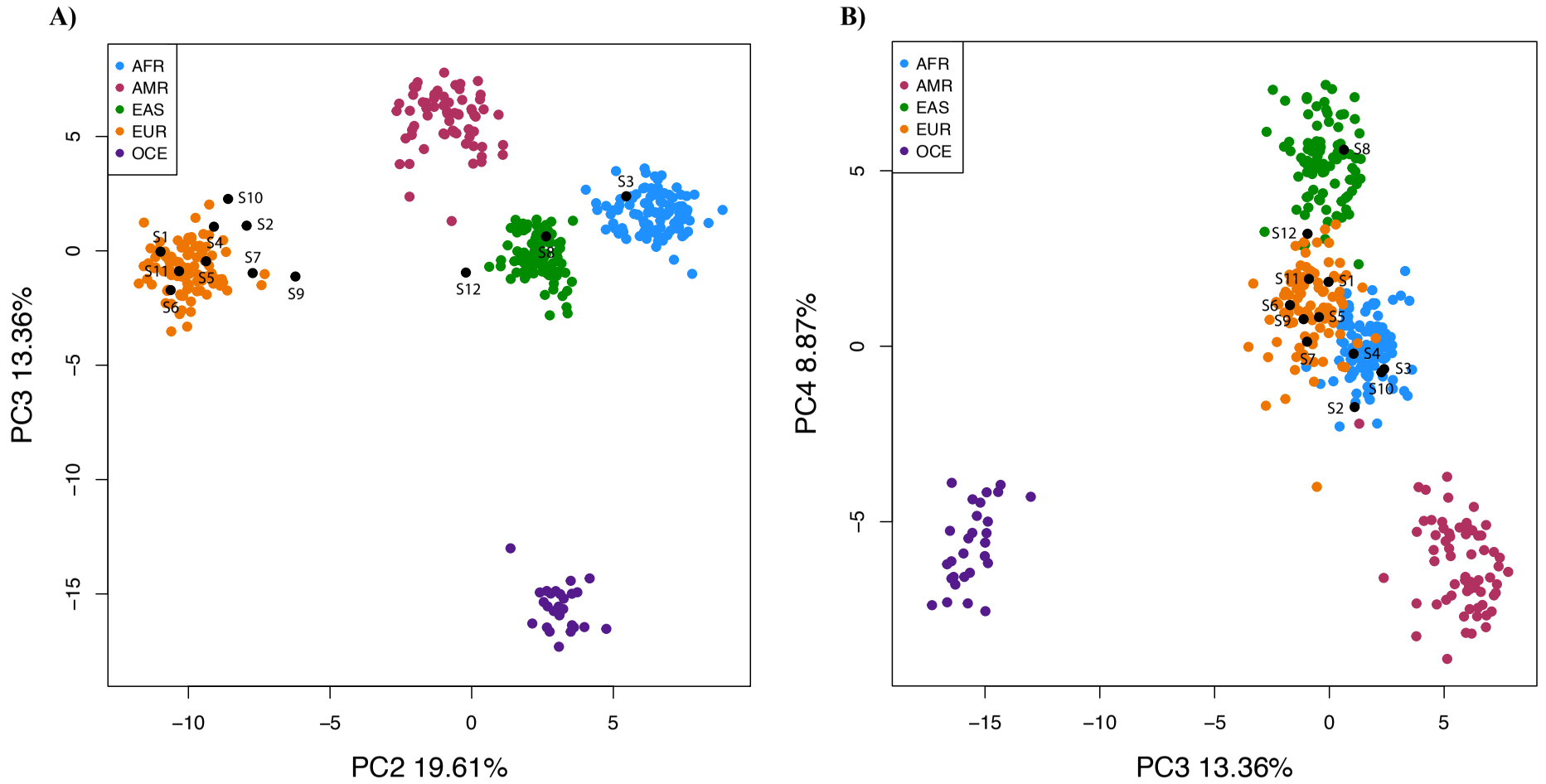
SNP	Mean Depth of Coverage	Standard Deviation
rs2032658	81	25
rs13447352	93	35
rs9341301	187	97
rs2032636	191	102
rs8179021	193	118
rs2032631	200	99
rs35284970	201	108
rs2032666	201	115
rs17250535	204	105
rs3900	206	109
rs3848982	209	108
rs9786025	214	127
rs9306841	224	140
rs11575897	230	137
rs2032595	234	139
rs13447371	236	138
rs9341308	236	133
rs868363758	239	140
rs2032668	240	144
rs371443469	243	137

SNP	Mean Depth of Coverage	Standard Deviation
rs17250163	245	148
rs9786153	246	144
rs13447354	249	146
rs2033003	256	151
rs1115657 (M)	260	157
rs9341278	264	142
rs13447361	271	169
rs9786706	274	161
rs3902	278	129
rs796808804	280	156
rs2032673	280	163
rs2032602	285	169
rs9341313	287	160
P308	288	164
rs9786140	301	185
rs9786184	316	211
rs4892491 (M)	344	189
rs1115657 (F)	664	241
rs4892491 (F)	793	254

Supplementary File S8. Intermediate likelihood ratios of ancestry predictions from Snipper for twelve samples with known ancestry

Sample	Self-declared ancestry	Region	Snipper Inferred Ancestry	Intermediate Likelihoods from Snipper
S5	European	Western Europe	European	2.1E+50 times more likely to be EUR than EAS and 1.5E+59 times more likely to be EUR than AFR
S12	East Asian	South East Asia	East Asian	1.4E+28 times more likely to be EAS than OCE and 1.4E+29 times more likely to be EAS than EUR
S2	Native American	Central America	European	7.6E+51 times more likely to be EUR than EAS and 9.3E+71 times more likely to be EUR than AFR
S3	African	Sub-Saharan Africa	African	1.8E+68 times more likely to be AFR than AMR and 2.5E+75 times likely to be AFR than EAS
S6	European	Western Europe	European	1.4E+58 times more likely to be EUR than EAS and 2.3E+58 times more likely to be EUR than AFR
S9	African	North Africa	European	6.6E+43 times more likely to be EUR than AMR and 7.0E+48 times more likely to be EUR than EAS
S1	European	Western Europe	European	3.6E+55 times more likely to be EUR than EAS and 4.3E+57 times more likely EUR than AFR
S4	European	Western Europe	European	4.4E+45 times more likely to be EUR than AFR and 3.4E+52 times more likely to be EUR than EAS
S7	European	Western Europe	European	2.0E+42 times more likely to be EUR than AFR and 1.4E+47 times more likely to be EUR than EAS
S8	East Asian	Mainland East Asia	East Asian	5.3E+44 times more likely to be EAS than OCE and 2.5E+60 times more likely to be EAS than EUR
S10	Native American	Central America	European	1.7E+50 times more likely to be EUR than AFR and 2.9E+51 times more likely to be EUR than EAS
S11	European	Western Europe	European	1.4E+56 times more likely to be EUR than EAS and 4.1E+58 times more likely to be EUR than OCE

Supplementary File S9. (A) PC2 versus PC3, and (B) PC3 versus PC4 of twelve study samples (black) with known ancestry against 368 genotypes across five global reference population groups. AFR (blue), AMR (magenta), EAS (green), EUR (orange), OCE (purple).



Supplementary File S10. Details of the 209 admixed population samples from the 1000 Genomes and HGDP-CEPH datasets used for comparison. All genotypes are presented in forward orientation (provided as an electronic file on USB drive).

Chapter 4

Application of the Miniplex SNaPshot assay and the 124-SNP hybridisation enrichment assay to degraded human DNA

Manuscript prepared for submission

Application of the Miniplex SNaPshot assay and the 124-SNP hybridisation enrichment assay to degraded human DNA

Felicia Bardan[1], Denice Higgins[1], Jeremy J. Austin[1]

[1]Australian Centre for Ancient DNA, School of Biological Sciences, University of Adelaide, Adelaide, South Australia 5005, Australia

*corresponding author: felicia.bardan@adelaide.edu.au

Statement of Authorship

Title of Paper	Application of the Miniplex SNaPshot assay and the 124-SNP hybridisation enrichment assay to degraded human DNA
Publication Status	☐ Published ☐ Accepted for Publication ☐ Submitted for Publication ☒ Unpublished and Unsubmitted work written in manuscript style
Publication Details	Manuscript prepared for publication

Principal Author

Name of Principal Author (Candidate)	Felicia Bardan
Contribution to the Paper	Helped conceive the study, helped generate, analyse and interpret the data, drafted the manuscript and produced the figures.
Overall percentage (%)	65%
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.
Signature	Date 22/10/18

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

i. the candidate's stated contribution to the publication is accurate (as detailed above);

ii. permission is granted for the candidate in include the publication in the thesis; and

iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author	Denice Higgins
Contribution to the Paper	Helped conceive the study, helped collect the samples, helped generate and analyse data, revised the manuscript
Signature	Date 22/10/18

Name of Co-Author	Jeremy J Austin
Contribution to the Paper	Helped conceive the study, helped collect the samples, and revised the manuscript.
Signature	Date 23/10/18

Abstract

Environmentally compromised and degraded samples can often fail to produce usable results with routine forensic genetic testing, resulting in partial or uninformative profiles. Recent developments in massively parallel sequencing of PCR multiplexes and hybridisation enrichment techniques offer solutions for typing hundreds of markers from degraded samples, however determining when a sample should be subjected to such techniques is usually at the cost of precious sample and resources. We recently developed a SNP-based screening and triaging tool, the 'Miniplex', and a custom hybridisation enrichment approach for use on generating intelligence data from degraded samples. In this study, we demonstrate the application of these methods to 38 degraded and forensic casework samples with post-mortem intervals of up to 70 years. The Miniplex provided an indication of sample degradation through comparison of mitochondrial and nuclear SNP typing success and allowed broad inferences of biological profile. Our custom enrichment panel was able to generate fine-resolution inferences for ancestry, hair and eye colour, sex and Y chromosome lineage to higher statistical likelihoods. Intelligence information gained from these methods are useful both in the selection and prioritisation of probative samples for downstream analysis, and for use in guiding forensic investigations of degraded remains where no other biological information can be gathered from a sample.

Keywords: degraded DNA; SNPs; biogeographic ancestry; phenotype; hybridisation enrichment

1. Introduction

Forensic casework samples for disaster victim and missing persons identification often contain limited amounts of highly degraded DNA (Higgins *et al.* 2015). This can lead to partial or uninformative DNA profiles when using conventional short tandem repeat (STR) markers (Higgins *et al.* 2015). Autosomal STR profiling is sensitive down to 62pg of DNA (Cornelis *et al.* 2018), or approximately 10 diploid cells. However the optimal range for most commercial kits is considered to be between 0.5 and 2ng of template DNA (or approximately 84-334 diploid cells) (Kline *et al.* 2005), which may not be available in degraded samples. Advances in post-extraction DNA analysis of degraded remains such as single nucleotide polymorphism (SNP) typing and massively parallel sequencing (MPS) are providing new ways to retrieve genetic information from the small amounts of highly degraded DNA recovered from such samples (Musgrave-Brown *et al.* 2007; Phillips *et al.* 2009; Fondevila *et al.* 2013; Templeton *et al.* 2013; Daniel *et al.* 2015; Gettings *et al.* 2015). Recent developments in MPS of polymerase chain reaction (PCR) multiplexes and alternative target enrichment techniques offer solutions to the genetic profiling of degraded remains, via simultaneous sequence analysis of hundreds of genetic markers (Churchill *et al.* 2016; de la Puente *et al.* 2017; Al-Asfi *et al.* 2018; Bose *et al.* 2018). However, resource, sample availability and specialised expertise are just some of the constraints of these technologies, and such tools remain impractical to implement for forensic laboratories with low-throughput MPS requirements. Currently available commercial MPS kits can also encounter difficulties with degraded DNA due to the amplicon sizes (>200 bp in some kits) (Gettings *et al.* 2015) and loss of intact primer binding sites. Careful decision-making to determine when samples should be subjected to more specialised techniques such as MPS is a concern for forensic laboratories.

Hybridisation enrichment-MPS technologies using short biotinylated probes have been explored recently for forensic identification purposes as an alternative approach to PCR-based multiplex enrichment (Templeton *et al.* 2013; Bose *et al.* 2018). One advantage of using a hybridisation capture approach is in the ability to retrieve very short DNA fragments – down to 30bp. Templeton *et al.* (2013) demonstrated an in-solution hybridisation enrichment technique that recovered whole mitochondrial genomes from forensically challenging post-mortem human skeletal remains ranging from 10 to ~2,500 years old. Bose *et al.* (2013) applied a custom hybridisation enrichment panel (307 SNPs and 36 microhaplotypes), with a focus on mixture detection using identity informative SNPs (Bose *et*

al. 2018). Both studies demonstrate the application of a hybridisation enrichment strategy for forensics purposes.

We recently developed an efficient and economical SNP-based triaging method (Miniplex) that provides a broad biological profile from mitochondrial DNA (mtDNA), Y chromosome (Y-chr) SNPs, and autosomal SNPs. The Miniplex also assesses sample quality by comparison of mtDNA and nuclear SNP recovery with varying amplicons lengths (66-128bp) (Chapter 2; Bardan *et al.* 2018). The Miniplex allows triaging of samples to determine the most suitable downstream identification workflow (i.e. STR genotyping or MPS techniques). Subsequently we developed a customisable hybridisation enrichment nuclear SNP panel to provide more detailed biogeographic ancestry, phenotype and Y-chr lineage (Chapter 3) and tested this on a set of modern, high quality human DNA samples with known ancestry, phenotype and sex. The panel correctly inferred biogeographic ancestry, hair and eye colour, Y-chr lineage and sex on most samples, indicating its value as a means to gather forensic intelligence from an unknown DNA sample. The objective of the current study is to apply and compare the Miniplex and the custom hybridisation enrichment panel to retrieve SNP-profiles and infer ancestry, phenotype, Y-chr haplogroup and sex from a range of degraded DNA and casework samples. We demonstrate the successful application of these novel methods to provide intelligence data for cases involving missing persons, and for degraded and skeletal human remains.

2. Materials and Methods

2.1 Degraded DNA technique and quality control

All steps preceding multiplex PCR and DNA library amplification were performed in dedicated low-copy and ancient DNA laboratories (geographically separate from post-amplification and modern molecular biology laboratories) to maintain integrity of DNA samples and mitigate the risk for contamination of samples with modern human DNA or PCR products. Stringent measures to minimise laboratory contamination were applied, including use of UV lights in all rooms and in all glove boxes, positive HEPA-filtered air pressure in laboratories, cleaning of work areas with sodium hypochlorite and isopropanol before and after use, personal protective equipment and triple-gloving during sample handling. Extraction blanks were included in each extraction batch. No-template controls were included during SNaPshot PCR set-up, library preparation and hybridisation enrichment. All controls were included to monitor potential contamination from human DNA sources and cross-contamination from other samples.

2.2 Samples and DNA extraction

In total, 38 samples were used including degraded teeth from a previous study (Higgins *et al.* 2015) and casework samples (human bone and hair). Sample details are given in Table 1.

2.2.1 Degraded Teeth

Thirty human teeth (14 male, 16 female) from a range of post-mortem intervals (1, 2, 4, 8, and 16 months) and previously extracted by Higgins *et al.* (2015) were used in accordance with ethics approval from the University of Adelaide Human Research and Ethics Committee (H-2016-198). Self-declared ancestry and phenotype (hair and eye colour) were not recorded for the donors.

2.2.2 Casework samples

Eight casework samples that had previously been extracted within the last ten years included six bones, one anagen hair and one hair shaft. The bones represent archaeological remains from Europe (1 sample), and unidentified remains from tropical and sub-tropical environments in Australia (4 samples, < 20 years old) and Papua New Guinea (1 sample, ~ 70 years old) and were mostly fragmentary and recovered from soil environments. The hair samples were recovered from a plaster bust of the "Somerton Man", an unidentified man of approximately 45 years of age found deceased on Somerton Beach, Adelaide, Australia in December 1948. His identity has never been established, but a death mask and subsequently a plaster bust was made of his head and shoulders in 1949 (5 months after death, (Adelaide News 1949)). The bust contains a number of hairs that are believed to come from the Somerton Man's head.

To reduce surface contamination, the outer surfaces of the bones were UV irradiated (260nm) for 30 mins, then ~1 mm of the sample surfaces was removed using a Dremel tool with a carborundum cutting disc. Each sample was then ground to a fine powder using a Mikro-Dismembrator (Sartorius). DNA was extracted from 0.2-0.5 g of powdered bone using a silica in-solution method (Brotherton *et al.* 2013) in a dedicated ancient DNA laboratory. Sex, ancestry and phenotype data were not known for bone samples. The hair samples were extracted using the Charge Switch Forensic DNA Purification Kit with a slight modification to the manufacturer's instruction (Edson *et al.* 2013).

Table 1. DNA samples used in this study. 'UNK' denotes samples for which sex was not known.

Sample No.	Sex	Sample Type	Sample Description
1	Male	Degraded Teeth	Tooth
2	Female	Degraded Teeth	Tooth
3	Female	Degraded Teeth	Tooth
4	Male	Degraded Teeth	Tooth
5	Male	Degraded Teeth	Tooth
6	Male	Degraded Teeth	Tooth
7	Female	Degraded Teeth	Tooth
8	Female	Degraded Teeth	Tooth
9	Female	Degraded Teeth	Tooth
10	Female	Degraded Teeth	Tooth
11	Male	Degraded Teeth	Tooth
12	Male	Degraded Teeth	Tooth
13	Female	Degraded Teeth	Tooth
14	Female	Degraded Teeth	Tooth
15	Female	Degrade Teeth	Tooth
16	Female	Degraded Teeth	Tooth
17	Female	Degraded Teeth	Tooth
18	Female	Degraded Teeth	Tooth
19	Male	Degraded Teeth	Tooth

Sample No.	Sex	Sample Type	Sample Description
20	Male	Degraded Teeth	Tooth
21	Male	Degraded Teeth	Tooth
22	Male	Degraded Teeth	Tooth
23	Male	Degraded Teeth	Tooth
24	Male	Degraded Teeth	Tooth
25	Female	Degraded Teeth	Tooth
26	Male	Degraded Teeth	Tooth
27	Female	Degraded Teeth	Tooth
28	Female	Degraded Teeth	Tooth
29	Female	Degraded Teeth	Tooth
30	Male	Degraded Teeth	Tooth
31	UNK	Casework	Bone
32	UNK	Casework	Bone
33	UNK	Casework	Bone
34	UNK	Casework	Bone
35	UNK	Casework	Bone
36	UNK	Casework	Bone
37	Male	Casework	Anagen hair
38	Male	Casework	Hair shaft

2.3 Miniplex PCR and SBE typing

SNaPshot SNP typing using the Miniplex PCR and SBE reactions were performed as described in (Bardan *et al.* 2018) (Chapter 2).

2.4 124-SNP hybridisation enrichment and massively paralleled sequencing

Library preparation and hybridisation enrichment were performed as described previously (Chapter 3) with the following alterations: the number of cycles determined by real-time PCR for first library amplification was capped at 15 cycles (regardless if results indicated more cycles were needed) in order to minimise clonality before enrichment. Samples that returned no detectable DNA on the Bioanalyzer 2100 (Agilent Technologies) following the second round of enrichment were subjected to real-time PCR using SYBR Green chemistry to determine the number of re-amplification cycles to generate sufficient material for sequencing (Chapter 3 Figure 1). DNA libraries were diluted 1:5 in DNA-free water and 1 uL was added to a final reaction volume of 10 uL comprising 1x HiFi buffer, 2 mM MgSO4, 250μM of each dNTP, 0.4 μM of each IS7 and IS8 primer (Meyer & Kircher (2010), 0.04 U Platinum HiFi Taq, and 0.4 uL ROX/SYBR mix (1 uL SYBR Green DNA Stain; Life Technologies, 4uL ROX; Thermo Fisher, 2 mL DMSO; Sigma). Thermocycling consisted of

a 6 min denaturation step at 94 °C, followed by 40 cycles of 15 s at 94°C, 15 s at 60°C, and 30 s at 68°C. All samples were run in duplicate and negative (PCR blank) controls were included on all runs. Real-time PCR was performed on a Roche LightCycler 96 thermocycler (Roche Life Sciences). Average Cq values generated for each of the samples indicated appropriate number of PCR cycles for reamplification.

Enriched samples were split into three runs and pooled equimolar to 5nM before sequencing on an Illumina MiSeq at the Australian Genome Research Facility (AGRF) using paired end 150 bp sequencing. For samples with low DNA quantities that did not meet the 5nM threshold, the total volume of enriched product was added into the pool. No DNA was detected in enriched negative controls using a Bioanalyzer but were spiked into the runs at a 10% volume of the final pool.

2.5 Data Analysis

Miniplex profile interpretation was performed as described in Bardan *et al*. 2018 (Chapter 2). In addition, STRUCTURE analysis (Porras-Hurtado *et al.* 2013) was performed on the Miniplex ancestry SNP profiles for comparison to MPS biogeographic ancestry data.

MPS data analysis and SNP calling were performed as previously described in Chapter 3 to generate a maximum 124-SNP (for males) or 89-SNP (for females) genotype for each sample. PCR duplicates (reads starting and ending at the same genome coordinates) were discarded so that only unique reads were included for genotype calling. For the inference of biogeographic ancestry, phenotype and Y-chr haplogroup, predictions were made using SNP calls at three different read depth thresholds (total locus coverage of ≥2x, ≥5x, and≥10x). Concordance for genotypes across the Miniplex and 124-plex hybridisation enrichment methods was checked for the 17 SNPs in common. Reference population data for the inference of biogeographic ancestry using the Miniplex and 124-plex hybridisation enrichment data are given in Supplementary File S1 and S2.

For hair and eye colour predictions using the HIrisPlex model, samples missing *HERC2* rs12913832 do not produce an eye colour prediction result. According to the predictive model, if all *MC1R* (10 SNPs) were not retrieved, no hair colour prediction could be made. If samples were missing *HERC2-SLC45A2-IRF4* loci, no hair or eye colour prediction could be made (Walsh *et al.* 2014).

3. Results

3.1 SNP typing success

3.1.1 Miniplex SNP recovery

DNA extract input for Miniplex PCR amplification ranged from <0.5 ng – 36.2 ng (Supplementary Table S3). Complete profiles (13 SNPs for females, 18 SNPs for males) were obtained from 12/38 samples (32%). Full mtDNA profiles (five SNPs) were retrieved from 37/38 samples, and full nuclear SNP profiles (8 for females, 13 for males) were obtained from 12/38 samples. For the 26 samples that produced partial nuclear profiles, SNP typing success ranged from 0 - 92% (average of 53%) (Figure 1). Sample 23 did not produce any nuclear SNPs but gave a full mtDNA 5-SNP profile. Sample 38 (hair shaft) produced no SNPs.

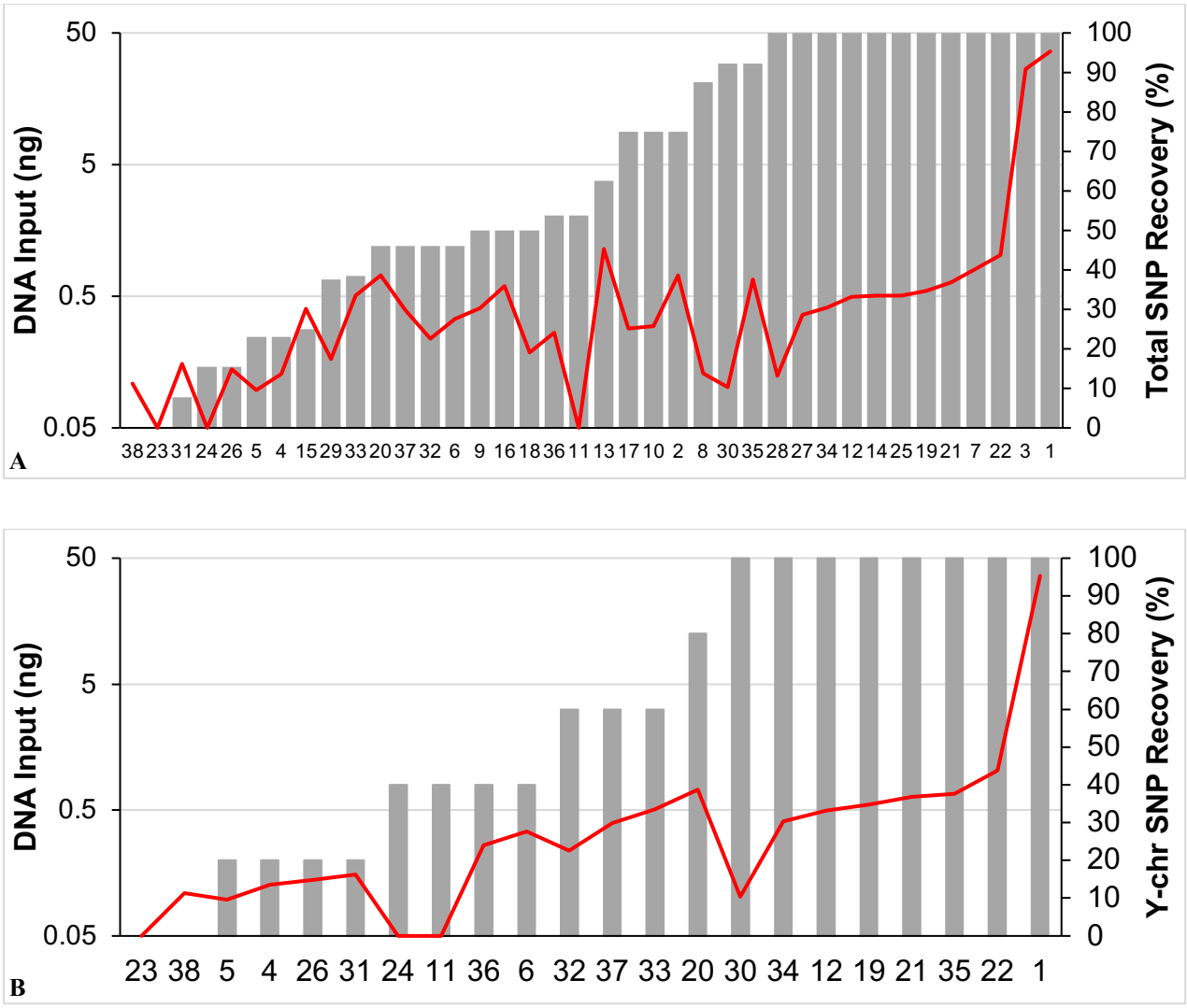


Figure 1. Total nuclear SNP typing success (A) and Y-chr SNP typing success (B) (grey bars) using the Miniplex on a range of degraded human teeth and casework samples with varying DNA input amounts. MtDNA SNP success not shown due to all but one sample returning complete mtDNA SNP profiles. Samples are ordered based on increasing SNP typing success in each graph. Details can be found in Supplementary Table S4.

3.1.2 Hybridisation enrichment performance and SNP recovery

Input DNA for hybridisation enrichment following library preparation ranged from 6.2 - 546 ng (Supplementary Table S3). The total number of reads retained per sample after quality filtering (retained reads) ranged from 2 to 1,904,915 (Table 2). Average read depth across the 38 samples for autosomal and diploid X-chr SNPs (female samples) was between 0 to 585, and approximately half this (0 to 241) for Y-chr and haploid X-chr SNPs (male). Samples with unknown sex ('UNK') *a piori* were considered to have haploid X-chr SNPs if they retrieved Y-chr SNPs. No Y-chr SNPs were retrieved from any of the known female samples. Clonality (the proportion of duplicate human sequences) was generally high, ranging from 19.7 – 99.5%, with a general trend towards lower clonality in samples with higher average read depth. Two samples (31 and 36) did not produce any of the target SNPs. Negative

controls did not obtain any retained or mapped reads and did not recover any of the targeted SNP markers (not shown).

Table 2. Number of retained reads, mapped reads, unique mapped reads, clonality and average read depth of coverage for autosomal (+ two diploid X chromosome SNPs for females) and Y-chromosome SNP loci (+ two haploid X chromosome SNPs for males) for 38 degraded teeth and casework samples. 'NA' denotes known female samples for which no reads for Y-chr SNPs were obtained.

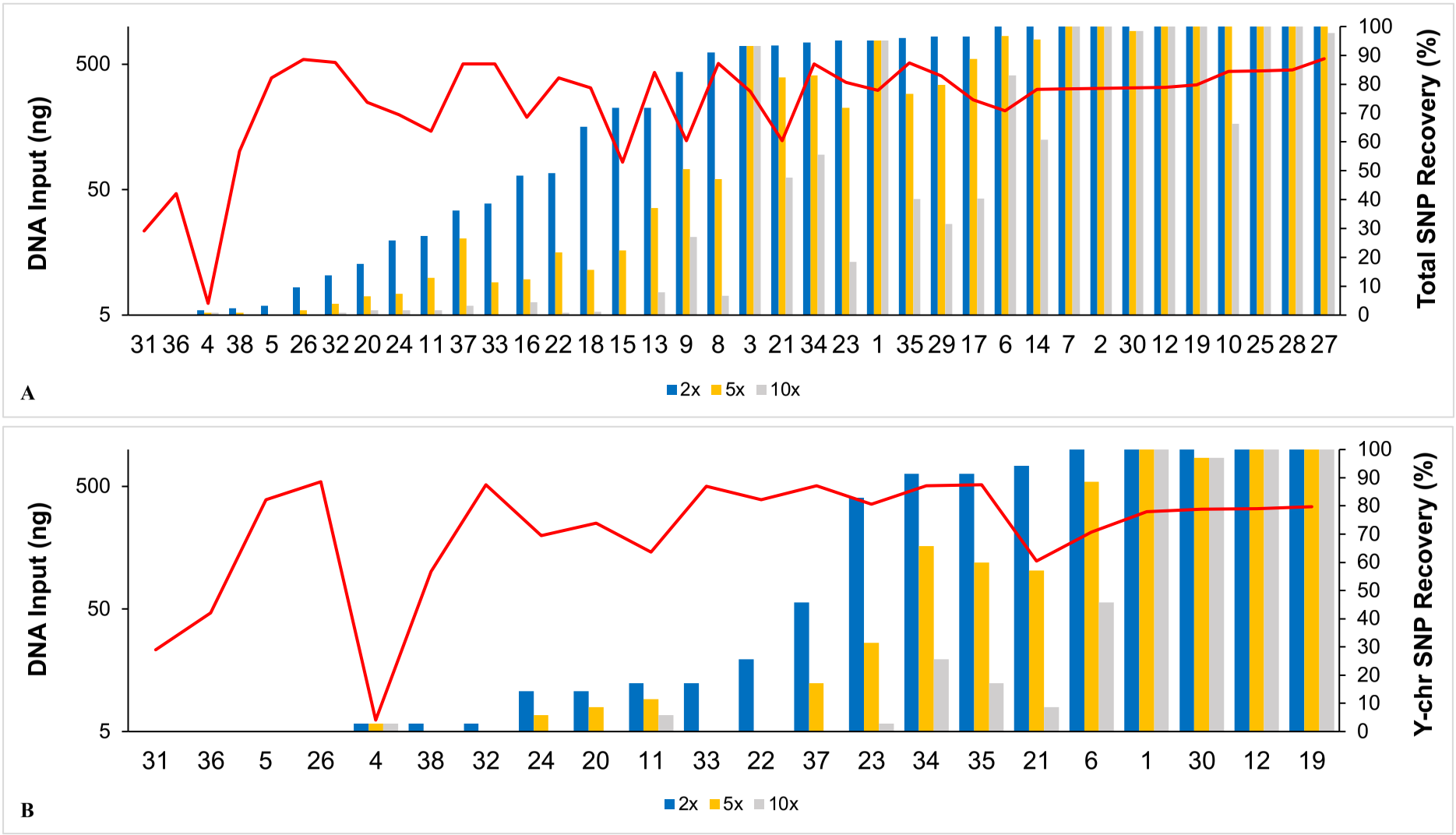
Sample	Sex	Retained Reads	Mapped Reads	Unique mapped reads	Clonality (%)	Average Depth (Autosomes + diploid X)	Average Depth (Y-chr + haploid X)
1	Male	180344	164450	120408	26.8	564	241
2	Female	543636	462811	120321	74.0	82	NA
3	Female	163295	148548	119336	19.7	585	NA
4	Male	62961	17914	4094	77.1	3	34
5	Male	10544	2227	135	93.9	2	0
6	Male	412670	326668	68896	78.9	22	9.0
7	Female	1321048	1147981	220295	80.8	51	NA
8	Female	976466	737127	102031	86.2	5	NA
9	Female	1084726	628121	12404	98.0	8	NA
10	Female	438381	289810	43526	85.0	12	NA
11	Male	1156237	1013512	54979	94.6	5	7
12	Male	491185	427187	108231	74.7	54	23
13	Female	1224478	731408	3439	99.5	6	NA
14	Female	1038238	773423	81427	89.5	12	NA
15	Female	170440	101801	10492	89.7	4	NA
16	Female	1268917	654976	16600	97.4	5	NA
17	Female	1250765	791664	10079	98.7	10	NA
18	Female	235910	194116	32141	83.4	4	NA
19	Male	1366779	1182315	253185	78.6	76	35
20	Male	1131004	547204	4309	99.2	5	4
21	Male	233039	157227	28831	81.7	11	6
22	Male	1477707	730051	4185	99.4	5	2
23	Male	479457	402428	57448	85.7	8	5
24	Male	1328445	904034	36873	95.9	4	4
25	Female	1177241	1093374	380822	65.2	377	NA
26	Male	636406	391241	870	99.8	4	0
27	Female	1047660	864463	202962	76.5	20	NA
28	Female	1656883	1288566	235524	81.7	27	NA
29	Female	247679	229311	4758	97.9	8	NA
30	Male	1205566	1141889	87689	92.3	39	21
31	UNK	519	7	5	28.6	0	0
32	UNK	215448	183325	619	99.7	4	3
33	UNK	176742	81685	1082	98.7	4	2
34	UNK	278397	256556	2038	99.2	14	8
35	UNK	1904915	1736174	1893	99.9	12	7
36	UNK	2	0	0	NA	0	0
37	Male	269534	235433	50586	78.5	5	4
38	Male	99661	96211	170	99.8	5	3

SNP recovery success was calculated at three different depths of coverage (2x, 5x, 10x) for the total suite of SNPs (autosomal, X, Y SNPs; Figure 2A), and for Y-chr SNPs (Figure 2B). Eleven samples (29%) gave full profiles at a minimum of 2x, eight (21%) at a minimum of 5x and six (16%) at a minimum of 10x read depth over all SNPs.

Average total SNP recovery (total SNPs possible for females is 89 and 124 SNPs for males) for 2x, 5x and 10x thresholds across the 38 samples was 63.3%, 53.7% and 39.1%, and average Y-chr SNP recovery across 22 samples was 45.6%, 33.9%, and 23% respectively. SNP typing success for samples with unknown sex was calculated with 124 possible SNPs regardless of whether Y-chr SNPs were called.

Coverage statistics indicated that the amount of DNA input to the hybridisation enrichment was not correlated with SNP recovery success at any read depth of coverage threshold (R^2 = 0.11 for 2x, R^2 = 0.09 for 5x, R^2 = 0.05 for 10x).

Figure 2. Total SNP recovery (A) and Y-chr SNP recovery (B) of custom enrichment panel on 38 degraded human teeth and casework samples with varying amounts of DNA input into the hybridisation reactions (red line) (only samples for which male sex was known, or samples which retrieved Y-chr SNPs are presented in B). SNP typing success is reported for a minimum read depth of 2x (blue bars), 5x (orange bars), and 10x (grey bars). Samples are ordered by increasing SNP recovery at 2x read depth in each graph. Details can be found in Supplementary Table S4.



3.2 Lineage-SNPs and sex determination

3.2.1 Miniplex

All five mtDNA SNPs were obtained from all but one sample (38), which retrieved no mtDNA SNPs. Twenty-eight samples were assigned to macrohaplogroup R, four samples to M, three samples to N, and one sample each to L3 and D (Table 3). The mtDNA profiles for each sample made phylogenetic sense and indicated only one haplogroup each.

A full Y-chr SNP profile was obtained from eight samples - six from 16 known males, and two where sex was unknown *a priori.* The highest rate of dropout of Y-chr SNPs was observed for indel M175. A haplogroup could not be assigned to two samples that retrieved full Y-chr profiles (i.e. SNPs eliminated haplogroup D, E, C, R and O). Seven known male samples had insufficient Y-chr SNPs to be assigned to a haplogroup (Table 4). The remaining 12 samples were assigned into R-M412 (10 samples), C-M216 (one sample) and D-M174 (one sample), including five casework samples where sex was not known. The Y -chr profiles made phylogenetic sense and indicated only one haplogroup each.

No 5-SNP Y-chr profile was retrieved for the 14 known female samples. Six female samples retrieved a T allele for Y-chr marker M174 shown previously to amplify in female samples (Chapter 2). Based on the criteria of obtaining >2 Y-chr SNPs for indicating male sex, 16 samples were determined as male, including five unknown samples. All known female samples were correctly predicted.

Table 3. mtDNA haplogroup results from 30 buried human teeth samples and eight casework samples using five mtDNA SNPs in the Miniplex.

Sample	mtDNA haplogroup
1	R
2	R
3	R
4	M
5	M
6	R
7	N
8	N
9	N
10	R
11	R
12	R
13	R
14	R
15	R
16	R
17	R
18	R
19	R

Sample	mtDNA haplogroup
20	R
21	R
22	M
23	M
24	L3
25	R
26	R
27	R
28	R
29	R
30	R
31	R
32	R
33	D
34	R
35	R
36	R
37	R
38	Could not classify

3.2.2 Hybridisation enrichment Y-chr SNPs

Full Y-chr SNP profiles (35 SNPs) were retrieved for 5/22 samples (23%) where either male sex was known (16 samples), or where sex was unknown (6 samples) *a priori*. No Y-chr SNPs were retrieved from any of the known female samples. A Y-chr haplogroup was predicted for twelve samples (Table 4). Ten samples (six of which were known to be male) did not retrieve sufficient diagnostic SNPs to allow for haplogroup assignment at a 2x read depth. Two samples which were known to be male (5 and 26), and two samples for which sex was unknown (31 and 36), did not retrieve any Y-chr SNPs.

For samples where a Y-chr haplogroup could be assigned (12 samples), all Y-chr SNPs made phylogenetic sense (e.g., no conflicting haplogroup assignments) across all three read depth thresholds. Two samples (11 and 22) only retrieved sufficient SNPs to assign a broad haplogroup deep rooted in the Y-chr phylogeny (KLT-M9, CDEF-M168). Seven samples were assigned to Y-chr haplogroup R-M412, one sample each to haplogroup R-M343, basal haplogroup G-M201, and haplogroup I-M258.

3.2.3 Comparison of the two methods

Four Y-chr SNPs targeted in the Miniplex were also targeted in the enrichment/MPS panel (herein referred to as 'enrichment method') (INDEL M175 was excluded from probe design). For ten samples able to be assigned a haplogroup using both genotyping methods, no

conflicting haplogroups were observed. The Miniplex was able to assign a broad haplogroup to four samples where the enrichment method could not. The enrichment method was able to assign a haplogroup to two samples where the Miniplex could not. Five known male samples, and one sample where sex was unknown, could not be assigned sex due a lack of autosomal and Y-chr SNP data using either method. SNP profiles were concordant for 37/38 samples between the Miniplex and enrichment method. In one sample (sample 20), the Miniplex obtained an 'A' genotype for Y-chr locus M412 (diagnostic for R), where the enrichment method retrieved a 'G' at 5x read depth. The Y-chr haplogroup predicted using the Miniplex was R, but the enrichment method did not retrieve sufficient diagnostic SNPs to classify a Y-chr haplogroup. Two samples predicted as 'Not D, E, C, R, O' using the Miniplex were resolved to the deep-rooted CDEF-M168 haplogroup (sample 22), and to haplogroup G-M201 (sample 30). Using the enrichment method, eight samples were able to be assigned to a haplogroup with a higher resolution than the Miniplex (e.g. sample 6). Six samples could not be classified into a Y-chr haplogroup using either method.

Table 4. Y-chr results from 22 degraded samples where male sex was known, or where sex was unknown ('UNK') using the Miniplex and the enrichment method at 2x, 5x, and 10x read depth of coverage thresholds. 'Not D, E, C, R, O' denotes samples that did not fall into any of the Miniplex Y haplogroups. 'MP' in 'Coverage' column refers to Miniplex results where maximum number of Y-chr SNPs that can be obtained is five. Maximum number of Y-chr SNPs for enrichment panel is 35. 'UND' = undetermined.

Sample	Sex	Coverage	No. of SNPs	Inferred Y-chr haplogroup	Continental Affiliation
1	Male	MP	5	R (R-M412)	W Europe
		2x	35	R-M412	W Europe
		5x	35	R-M412	W Europe
		10x	35	R-M412	W Europe
4	Male	MP	1	Could not classify	-
		2x	1	Could not classify	-
		5x	1	Could not classify	-
		10x	1	Could not classify	-
5	Male	MP	1	Could not classify	-
		2x	0	Could not classify	-
		5x	0	Could not classify	-
		10x	0	Could not classify	-
6	Male	MP	2	R (R-M412)	W Europe
		2x	35	R-M412	W Europe
		5x	31	R-M412	W Europe
		10x	16	R-M412	W Europe
11	Male	MP	2	Could not classify	-
		2x	6	KLT-M9	UND
		5x	3	CDEF-M168	UND
		10x	2	CDEF-M168	UND
12	Male	MP	5	R (R-M412)	W Europe
		2x	35	R-M412	W Europe
		5x	35	R-M412	W Europe
		10x	35	R-M412	W Europe
19	Male	MP	5	R (R-M412)	W Europe
		2x	35	R-M412	W Europe
		5x	35	R-M412	W Europe
		10x	35	R-M412	W Europe
20	Male	MP	4	R (R-M412)	W Europe
		2x	5	Could not classify	-
		5x	3	Could not classify	-
		10x	2	Could not classify	-

Sample	Sex	Coverage	No. of SNPs	Inferred Y-chr haplogroup	Continental Affiliation
24	Male	MP	2	Could not classify	-
		2x	5	Could not classify	-
		5x	2	Could not classify	-
		10x	0	Could not classify	-
26	Male	MP	1	Could not classify	-
		2x	0	Could not classify	-
		5x	0	Could not classify	-
		10x	0	Could not classify	-
30	Male	MP	5	Not D, E, C, R, O	UND
		2x	35	G-M201	Middle East/Europe/Asia
		5x	34	G-M201	Middle East/Europe/Asia
		10x	34	G-M201	Middle East/Europe/Asia
31	UNK	MP	1	Could not classify	-
		2x	0	Could not classify	-
		5x	0	Could not classify	-
		10x	0	Could not classify	-
32	UNK	MP	3	R (R-M412)	W Europe
		2x	1	Could not classify	-
		5x	0	Could not classify	-
		10x	0	Could not classify	-
33	UNK	MP	3	D-M174	East Asia
		2x	6	Could not classify	-
		5x	0	Could not classify	-
		10x	0	Could not classify	-
34	UNK	MP	5	R (R-M412)	W Europe
		2x	32	R-M412	W Europe
		5x	23	R-M412	W Europe
		10x	9	R-M343	W Europe
35	UNK	MP	5	R (R-M412)	W Europe
		2x	33	R-M412	W Europe
		5x	21	R-M412	W Europe
		10x	6	R-M343	Europe/Africa

21	Male	MP	5	R (R-M412)	W Europe
		2x	33	R-M412	W Europe
		5x	20	R-M412	W Europe
		10x	3	R-M412	W Europe
22	Male	MP	5	Not D, E, C, R, O	UND
		2x	9	CDEF-M168	UND
		5x	0	Could not classify	-
		10x	0	Could not classify	-
23	Male	MP	0	Could not classify	-
		2x	29	I-M258	Europe
		5x	8	CDEF-M168	UND
		10x	1	Could not classify	-

36	UNK	MP	2	C-M216	Asia/Oceania/N America
		2x	0	Could not classify	-
		5x	0	Could not classify	-
		10x	0	Could not classify	-
37	Male	MP	3	R (R-M412)	Europe
		2x	16	R-M343	Europe/Africa
		5x	6	P-M45	Asia/America/Europe/Africa
		10x	0	Could not classify	
38	Male	MP	0	Could not classify	-
		2x	1	Could not classify	-
		5x	0	Could not classify	-
		10x	0	Could not classify	-

3.3 Biogeographic ancestry

3.3.1 Miniplex Snipper predictions

Of the 38 samples, 13 retrieved a full 5-SNP ancestry profile (34%). An ancestry prediction was possible in Snipper for 33 samples. Twenty-six samples were predicted as 'EUR', four samples returned an 'AFR' classification, and one sample as 'AMR' and 'EAS' each using the Bayesian classifier Snipper (Table 5). Sample 33 could not be classified into any one population, however AFR ancestry was excluded based on its *-log(LIKELIHOOD)* value (the *-log(LIKELIHOOD* values for AMR, EAS, EUR and OCE were too similar for classification into one population group) (Supplementary File S5).

Five samples could not be classified due to retrieving no autosomal ancestry SNPs and were removed from subsequent Miniplex ancestry analysis.

3.3.2 Enrichment method Snipper predictions

A full 67-SNP ancestry profile (including four tri-allelic SNPs) was obtained from 12 samples (32%) using a read depth threshold of 2x, with an average of 48 SNPs recovered (a 72% recovery rate). For the minimum of 5x and 10x read depths, full ancestry SNP profiles were obtained from 10 and 6 samples, with an average SNP recovery of 39 (58%) and 28 (42%) SNPs respectively. Two samples (31 and 36) produced no SNPs and were excluded from ancestry analysis.

Using a minimum read depth threshold of 2x, 30 of the remaining 36 samples were predicted as having 'EUR' ancestry, 3 were predicted as 'AMR' ancestry, one sample was predicted as having 'EAS' ancestry and one as 'OCE' (Table 5). One sample could not be classified into either AMR, EAS, EUR or OCE ancestry, but AFR ancestry was excluded based on the *-log(LIKELIHOOD)* value. Different ancestry inferences for six samples were obtained when using a 2x read depth threshold for SNP calling to a 10x read depth, however only 1-2 SNPs were available at 10x read depth for ancestry inference versus 4 – 46 SNPs at 2x.

Table 5. Inferred biogeographic ancestry from Snipper for 38 degraded teeth and casework samples using the Miniplex (MP) and enrichment method at 2x, 5x, and 10x read depth thresholds. The lowest likelihood ratio is presented as symbols. Exact likelihood ratios from lowest to highest are given in Supplementary Table S5). 'MP' in 'Coverage' column refers to Miniplex results where maximum number of ancestry SNPs that can be obtained is five. Maximum number of SNPs for enrichment method is 67.

Sample	Coverage	SNPs	Prediction
1	MP	5	EUR #
	2x	64	EUR *
	5x	64	EUR *
	10x	64	EUR *
2	MP	3	EUR ⍙
	2x	67	EUR *
	5x	67	EUR *
	10x	67	EUR *
3	MP	5	EUR ⍙
	2x	64	EUR *
	5x	64	EUR *
	10x	64	EUR *
4	MP	0	NA
	2x	1	NA
	5x	0	NA
	10x	0	NA
5	MP	2	AFR △
	2x	4	EUR △
	5x	0	NA
	10x	0	NA
6	MP	2	EUR ○
	2x	67	EUR *
	5x	67	EUR *
	10x	65	EUR *
7	MP	5	EUR #
	2x	67	EUR *
	5x	67	EUR *
	10x	67	EUR *
8	MP	4	AMR ☆
	2x	61	AMR □
	5x	33	AMR ⬠
	10x	5	AMR △
9	MP	2	EAS △
	2x	55	AMR *
	5x	29	AMR +
	10x	17	AMR ○
10	MP	4	EUR ○
	2x	67	EUR *
	5x	67	EUR *
	10x	40	EUR *
11	MP	3	EUR ○
	2x	23	EUR +
	5x	11	EUR ⬠
	10x	0	NA
12	MP	5	EUR #
	2x	67	EUR *
	5x	67	EUR *
	10x	67	EUR *
13	MP	2	EUR △
	2x	48	EUR *
	5x	28	EUR *
	10x	7	EUR ○
14	MP	5	EUR #
	2x	67	EUR *
	5x	64	EUR *
	10x	40	EUR *
15	MP	2	EUR
	2x	47	EUR *
	5x	13	EUR ⬠
	10x	0	NA
16	MP	1	AFR △
	2x	32	EUR *
	5x	9	EUR ◇
	10x	2	OCE △
17	MP	3	EUR ◇
	2x	65	EUR *
	5x	56	EUR *
	10x	24	EUR *
18	MP	2	EUR ○
	2x	46	EUR *
	5x	11	EUR #
	10x	1	AFR △
19	MP	5	EUR #
	2x	67	EUR *
	5x	67	EUR *
	10x	67	EUR *
20	MP	2	AFR △
	2x	14	EUR *
	5x	5	EUR ◇
	10x	2	AFR △
21	MP	5	EUR #
	2x	64	EUR *
	5x	63	EUR *
	10x	41	EUR *
22	MP	5	EUR ◇
	2x	38	EUR *
	5x	16	EUR ○
	10x	1	AFR △
23	MP	0	NA
	2x	67	AMR *
	5x	58	AMR *
	10x	20	AMR ⍙
24	MP	0	NA
	2x	19	EUR ◇
	5x	6	EAS △
	10x	2	EAS △
25	MP	5	EUR
	2x	67	EUR *
	5x	67	EUR *
	10x	67	EUR *
26	MP	1	EUR ◇
	2x	4	EUR ☆
	5x	2	AFR △
	10x	2	AFR △
27	MP	5	EUR #
	2x	67	EUR *
	5x	67	EUR *
	10x	65	EUR *
28	MP	5	EUR #
	2x	67	EUR *
	5x	67	EUR *
	10x	67	EUR *
29	MP	1	EUR ◇
	2x	65	EUR *
	5x	52	EUR *
	10x	18	EUR ⬠
30	MP	4	EUR ⍙
	2x	67	EUR *
	5x	67	EUR *
	10x	66	EUR *
31	MP	0	NA
	2x	0	NA
	5x	0	NA
	10x	0	NA
32	MP	2	AFR △
	2x	15	EUR *
	5x	5	EUR △
	10x	1	NA
33	MP	1	NA
	2x	34	EAS *
	5x	10	EAS ◇
	10x	0	NA
34	MP	5	EUR #
	2x	64	EUR *
	5x	60	EUR *
	10x	43	EUR *
35	MP	5	EUR #
	2x	66	EUR *
	5x	55	EUR *
	10x	28	EUR *
36	MP	3	EUR ○
	2x	0	NA
	5x	0	NA
	10x	0	NA
37	MP	2	EUR △
	2x	46	EUR *
	5x	17	EUR *
	10x	4	EUR ⍙
38	MP	0	NA
	2x	2	OCE △
	5x	1	NA
	10x	0	NA

Likelihood Ratio Legend:

△ = 1 – 10	# = 100,000 – 1,000,000
☆ = 10 – 100	□ = 1,000,000 – 10,000,000
◇ = 100 – 1000	⬠ = 10,000,000 – 100,000,000
○ = 1000 – 10,000	+ = 100,000,000 – 1,000,000,000
⍙ = 10,000 – 100,000	* = > 1,000,000,000

3.3.3 Comparison of the two methods

All five ancestry SNPs targeted in the Miniplex were also included in the enrichment method. The Miniplex could infer ancestry using Snipper for one sample where the enrichment method was not able to due to a lack of informative SNPs (sample 36). At a minimum threshold of 2x read depth, the enrichment method allowed an ancestry inference for three samples (sample 24, 33, 38) where the Miniplex could not. Ancestry was not able to be inferred for two samples using either method (sample 4 and 31) due to a lack of informative SNPs.

Genotypes were concordant for all samples between the Miniplex and enrichment method except for one locus in one sample. The Miniplex obtained a 'GG' genotype for rs9908046 (informative for Oceanian ancestry) in sample 2, where the enrichment method retrieved 'AG' at 80x read depth (54% of reads for dominant allele A). The resulting ancestry classification however was concordant with both samples classified as 'EUR' in Snipper, STRUCTURE and PCA. However, in all cases only 1-2 SNPs were recovered for the Miniplex (resulting in likelihood ratios of ancestry prediction of only 1.1-1.2) compared to 4-55 SNPs recovered for the enrichment method (resulting in likelihood ratios of ancestry prediction of 2.9 to > 1,000,000).

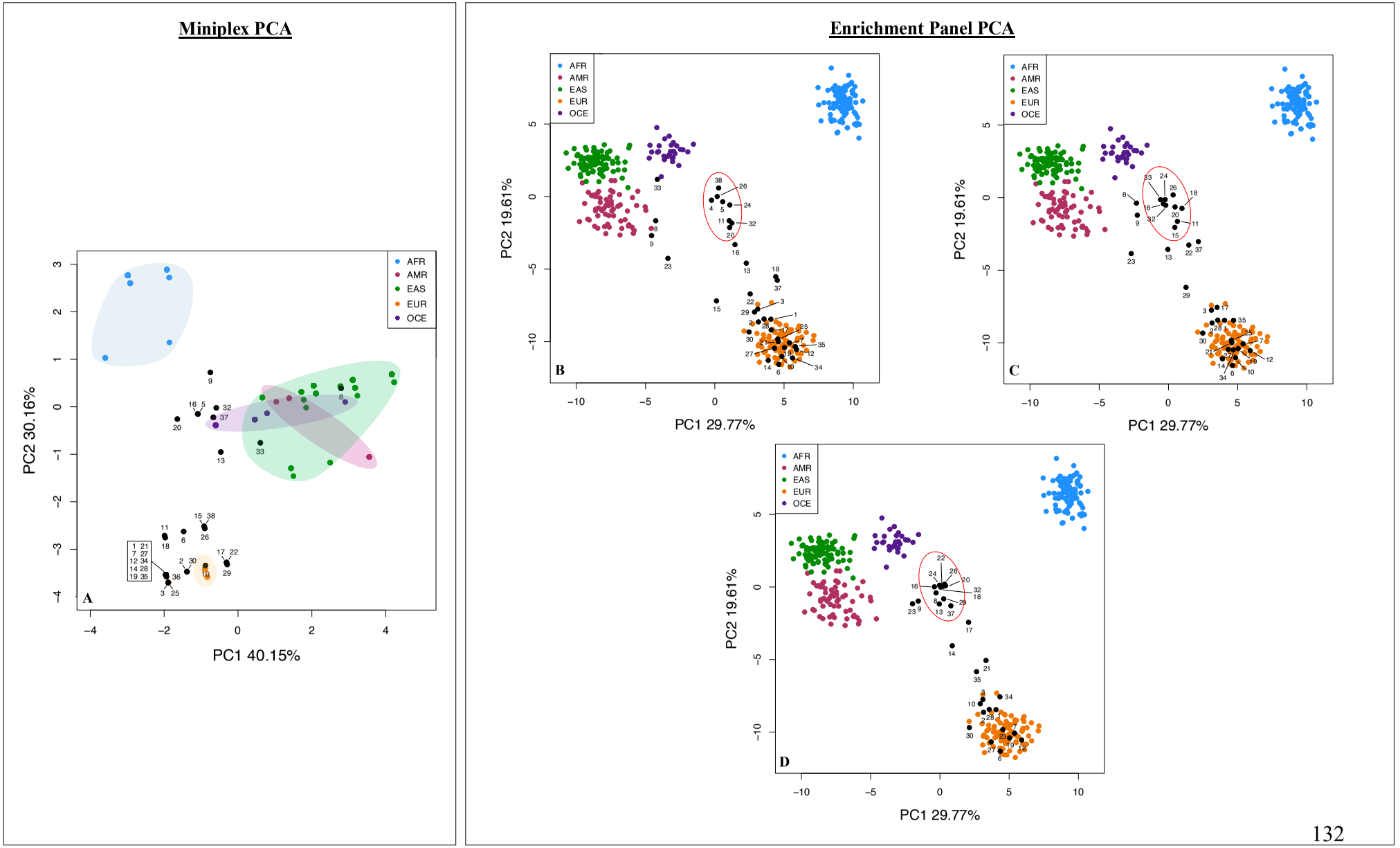
Table 6. Samples for which differing ancestry inferences were obtained from the Miniplex (MP) versus enrichment method at a minimum read depth of 2x. Maximum number of Miniplex SNPs is five. Maximum number of MPS SNPs is 67. 'LR' = likelihood ratio. Remaining LR values are given in Supplementary File S5.

Sample	MP SNPs	Prediction	Miniplex LR	MPS SNPs	Prediction	MPS LR
5	2	AFR	1 .1 more likely than EUR	4	EUR	2.9 more likely than AFR
9	2	EAS	1.1 more likely than AMR	55	AMR	> 1,000,000 more likely than EAS
16	1	AFR	1 .1 more likely than EUR	32	EUR	> 1,000,000 more likely than EAS
20	2	AFR	1 .1 more likely than EUR	14	EUR	> 1,000,000 more likely than EAS
32	2	AFR	1.2 more likely than EUR	15	EUR	> 1,000,000 more likely than AMR

The PCA plot for the five Miniplex ancestry SNPs from 34 study samples and 402 reference population genotypes shows the ancestry affiliation of the teeth and casework samples and are consistent with Snipper predictions (Figure 3A). Samples 4, 24, 31 and 38 were omitted from PCA since they retrieved no ancestry SNPs using the Miniplex. Results for the ancestry SNPs in the enrichment method (tri-allelic SNPs were excluded from PCA) using 368 reference population genotypes and 36 study samples (Figure 3B, 3C, and 3D) shows the ancestry affiliation of the teeth and casework samples and are broadly consistent with Snipper predictions obtained using the enrichment method. Samples 31 and 36 were omitted from

PCA since no ancestry SNPs were retrieved. Samples 4 and 5 were further omitted from PCA using 5x read depth, and samples 11, 15, 33, and 38 were further omitted from PCA using 10x read depth.

Figure 3. PCA plots (PC1 vs PC2) of degraded samples (black dots) using the Miniplex (A) and using the enrichment method at different minimum read depths (B=2x; C=5x; D=10x) against reference population genotypes detailed in Supplementary File S1 and S2. Single coloured dots in Fig.3A may represent >1 sample if they obtained the same SNP genotype across the 5 ancestry SNPs in the Miniplex. AFR (blue); EUR (orange); AMR (magenta); EAS (green) and OCE (purple). Red circles include samples for which ≤19 SNPs were obtained and didn't exhibit sufficient variation to separate into informative components.



Further analysis of the samples in STRUCTURE shows the ancestry components described using both the Miniplex (Figure 4B), and the enrichment method (Figure 4D). STRUCTURE analysis of the five Miniplex SNPs shows the separation of the five reference population groups giving clear distinction of AFR, EUR and OCE populations, but overlapping patterns between AMR and EAS population groups (Figure 4A). Samples 4, 23, 24, 31 and 38 were omitted since no ancestry SNPs were recovered using the Miniplex. Average STRUCTURE population membership coefficients showed 31 out of 33 samples had EUR ancestry as their major ancestry component, one had AMR as the major ancestry component, and one had EAS as the major ancestry component (ancestry proportions given in Supplementary Table S6).

STRUCTURE results for both reference population data and samples are broadly consistent with the PCA, and for Snipper predictions using the Miniplex, except for four samples. Sample 5, 16, 20 and 32 indicated a EUR major ancestry component in STRUCTURE, whereas Snipper predicted AFR ancestry for these four samples. STRUCTURE ancestry membership coefficients are detailed below in Table 7.

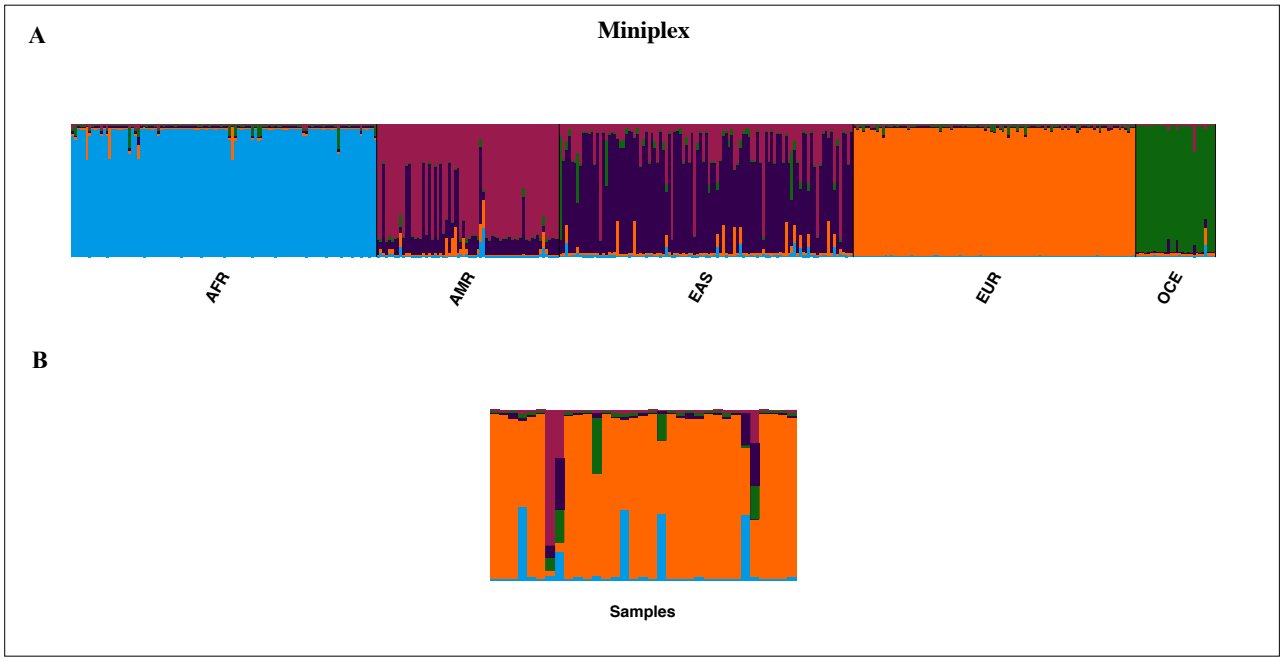
Table 7. Population membership proportions of three samples that were predicted as AFR in Snipper but showed a EUR major ancestry component in STRUCTURE.

Sample	Snipper prediction	% AFR	% AMR	% EAS	% EUR	% OCE
5	AFR	40.9	2.3	2.3	52.4	2.2
16	AFR	41.7	1.9	2.2	52.5	1.73
20	AFR	39.2	1.1	1.5	42.7	15.4
32	AFR	38.7	7.6	13.2	38.9	1.5

For the enrichment method, STRUCTURE analysis performed using 2x, 5x and 10x read depth SNP calling thresholds shows the ancestry components are broadly consistent with PCA and Snipper predictions. Samples 31 and 36 were omitted from analysis since no ancestry SNPs were retrieved at any of the read depth thresholds. Samples 4 and 5 were further omitted from CLUMPAK visualisation (grey bars) using a 5x read depth, and samples 11, 15, 33, and 38 were further omitted from CLUMPAK visualisation (grey bars) using a minimum of 10x read depth.

Average STRUCTURE population membership coefficients showed 31 out of 36 samples had EUR ancestry as their major ancestry component, three had AMR as their major ancestry component, and one each had EAS and OCE as their major ancestry component. One sample (sample 4), was not able to be predicted in Snipper but showed a major EUR ancestry

component at 31.7% in STRUCTURE. The remaining samples were consistent with Snipper predictions and PCA (ancestry proportions given in Supplementary Table S6).



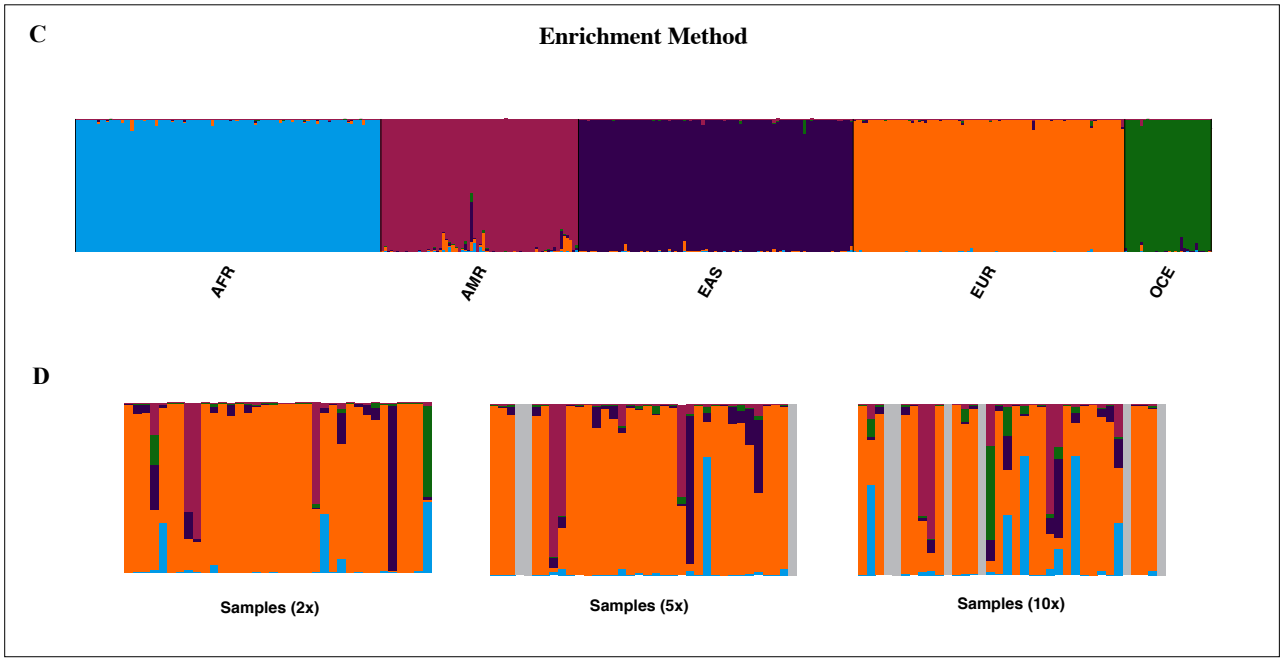


Figure 4. STRUCTURE analysis for reference population data and degraded study samples using the Miniplex (A & B) and enrichment method (C & D) at 2x, 5x and 10x read depth thresholds (K=5). Each vertical bar represents one individual. Samples are ordered according to Table 5. Samples 4, 23, 24, 31 and 38 were omitted from Miniplex analysis since no ancestry SNPs were recovered. Samples 31 and 36 were omitted from enrichment method analysis since no ancestry SNPs were recovered. AFR: African, AMR: Native America, EAS: East Asian, EUR: European, OCE: Oceanian. Grey bars indicate samples that did not retrieve any ancestry SNPs.

For male samples, comparison of Y-chr paternal ancestry (Table 4) and autosomal ancestry (Table 5) showed that continental ancestry predictions were consistent for all samples except sample 23. This sample, predicted as AMR in Snipper, was resolved into AMR and EUR ancestry components in STRUCTURE (AMR = 62.1%; EUR = 35.9%), but displayed a European Y-chr haplogroup using the enrichment method. These results indicate admixture between AMR and EUR ancestry for this sample.

3.4 Phenotype

3.4.1 Miniplex

All three phenotype SNPs were obtained from 18 samples (47%). Twenty-eight out of 38 samples retrieved sufficient SNPs to produce an eye colour prediction using the IrisPlex webtool. Of these 28 samples, 14 returned a 'not brown' eye colour result, and 14 were predicted as having brown eyes (Table 8).

Ten samples did not retrieve SNP rs12913832, and consequently an eye colour prediction was not able to be made. The phenotype SNPs in the Miniplex were not chosen to differentiate between hair colour, so hair colour predictions were not made.

3.4.2 Enrichment method

All 23 phenotype SNPs were obtained from twelve samples (32%) using the enrichment method. Twenty-six samples were able to produce both an eye and hair colour prediction. Eye colour predictions are given in Table 8, and hair colour predictions are given in Table 9.

Fourteen samples were predicted as having blue eyes, and the remaining twelve samples were predicted as having brown eyes. No intermediate eye phenotypes were predicted. Four samples (9, 13, 16, 22) could not be classified into an eye colour phenotype despite retrieving a moderate number of SNPs (21, 12, 14 and 16 SNPs out of 23 respectively) due to missing the rs12913832 SNP.

On the basis of the HIrisPlex step-wise model for inferring 'most probable hair colour', eleven samples were predicted as having brown/dark-brown hair, six as blond/dark-blond, two as dark-brown/black, two as black, two as red, and one as blond hair.

Twelve samples did not retrieve sufficient SNPs for both hair and eye colour prediction.

3.4.3 Comparison of the two methods

All three phenotype SNPs targeted in the Miniplex were also included in the enrichment method. When comparing predictions from the Miniplex and the enrichment method for eye colour, all but one was concordant. Despite having concordant genotypes between the same loci, sample 7 was predicted as having brown eyes using the Miniplex but was predicted as having blue eyes when using the 23 SNPs obtained from the enrichment method. However, the predictions from both the Miniplex and enrichment method retrieved a probability below the current reporting threshold of <0.7 despite retrieving a full profile for each, and so confident eye colour classification for this sample could not be made.

Table 8. Inferred eye colour for 38 samples, using the Miniplex data (with the IrisPlex webtool) and the enrichment method data (with the HIrisPlex webtool) at 2x, 5x, and 10x read depth thresholds. P-value scores are out of a value of 1, only the highest p-value is reported. 'MP' in 'Coverage' column refers to Miniplex results where maximum number of phenotype SNPs that can be obtained is three. Maximum number of phenotype SNPs for enrichment panel is 23. Remaining p-values and AUC loss values are given in Supplementary File S5.

Sample	Coverage	No. of SNPs	Inferred eye colour (p-value)
1	MP	3	Not Brown (0.917)
	2x	20	Blue (0.955)
	5x	20	Blue (0.955)
	10x	20	Blue (0.955)
2	MP	2	Not Brown (0.91)
	2x	23	Blue (0.891)
	5x	23	Blue (0.891)
	10x	23	Blue (0.891)
3	MP	3	Not Brown (0.892)
	2x	20	Blue (0.831)
	5x	20	Blue (0.831)
	10x	20	Blue (0.831)
4	MP	2	Brown (0.965)
	2x	0	Cannot classify
	5x	0	Cannot classify
	10x	0	Cannot classify
5	MP	0	Cannot classify
	2x	0	Cannot classify
	5x	0	Cannot classify
	10x	0	Cannot classify
6	MP	3	Not Brown (0.917)
	2x	23	Blue (0.942)
	5x	23	Blue (0.942)
	10x	23	Blue (0.942)
7	MP	3	Brown (0.448)
	2x	23	Blue (0.482)
	5x	23	Blue (0.482)
	10x	23	Blue (0.482)
8	MP	3	Brown (0.975)
	2x	21	Brown (0.998)
	5x	10	Brown (0.997)
	10x	2	Cannot classify
9	MP	2	Brown (0.965)
14	MP	3	Not Brown (0.917)
	2x	23	Blue (0.884)
	5x	23	Blue (0.884)
	10x	14	Blue (0.877)
15	MP	0	Cannot classify
	2x	17	Blue (0.852)
	5x	9	Cannot classify
	10x	0	Cannot classify
16	MP	2	Brown (0.965)
	2x	12	Cannot classify
	5x	2	Cannot classify
	10x	2	Cannot classify
17	MP	3	Brown (0.708)
	2x	22	Brown (0.668)
	5x	20	Brown (0.668)
	10x	12	Brown (0.501)
18	MP	2	Cannot classify
	2x	12	Brown (0.965)
	5x	4	Cannot classify
	10x	0	Cannot classify
19	MP	3	Not Brown (0.917)
	2x	23	Blue (0.911)
	5x	23	Blue (0.911)
	10x	23	Blue (0.911)
20	MP	0	Cannot classify
	2x	3	Cannot classify
	5x	0	Cannot classify
	10x	0	Cannot classify
21	MP	3	Not Brown (0.917)
	2x	20	Blue (0.889)
	5x	20	Blue (0.889)
	10x	18	Blue (0.927)
22	MP	3	Brown (0.977)
27	MP	3	Not Brown (0.917)
	2x	23	Blue (0.932)
	5x	23	Blue (0.932)
	10x	23	Blue (0.932)
28	MP	3	Brown (0.448)
	2x	23	Brown (0.512)
	5x	23	Brown (0.512)
	10x	22	Brown (0.512)
29	MP	1	Cannot classify
	2x	21	Blue (0.945)
	5x	18	Blue (0.895)
	10x	9	Cannot classify
30	MP	3	Not Brown (0.917)
	2x	23	Blue (0.903)
	5x	23	Blue (0.903)
	10x	23	Blue (0.903)
31	MP	0	Cannot classify
	2x	0	Cannot classify
	5x	0	Cannot classify
	10x	0	Cannot classify
32	MP	1	Not Brown (0.912)
	2x	0	Cannot classify
	5x	0	Cannot classify
	10x	0	Cannot classify
33	MP	1	Brown (0.969)
	2x	18	Brown (0.978)
	5x	5	Cannot classify
	10x	0	Cannot classify
34	MP	3	Not Brown (0.917)
	2x	22	Blue (0.959)
	5x	21	Blue (0.959)
	10x	18	Blue (0.957)
35	MP	2	Brown (0.66)

	2x	21	Cannot classify
	5x	16	Cannot classify
	10x	7	Cannot classify
10	MP	2	Not Brown (0.918)
	2x	23	Blue (0.848)
	5x	23	Blue (0.848)
	10x	19	Blue (0.84)
11	MP	3	Brown (0.975)
	2x	5	Brown (0.972)
	5x	2	Cannot classify
	10x	0	Cannot classify
12	MP	3	Brown (0.975)
	2x	23	Brown (0.974)
	5x	23	Brown (0.974)
	10x	23	Brown (0.974)
13	MP	3	Not Brown (0.917)
	2x	16	Cannot classify
	5x	7	Cannot classify
	10x	0	Cannot classify

	2x	14	Cannot classify
	5x	11	Cannot classify
	10x	1	Cannot classify
23	MP	0	Cannot classify
	2x	23	Brown (0.998)
	5x	20	Brown (0.998)
	10x	2	Cannot classify
24	MP	0	Cannot classify
	2x	8	Brown (0.948)
	5x	1	Cannot classify
	10x	0	Cannot classify
25	MP	3	Brown (0.708)
	2x	23	Brown (0.76)
	5x	23	Brown (0.76)
	10x	23	Brown (0.76)
26	MP	0	Cannot classify
	2x	8	Cannot classify
	5x	2	Cannot classify
	10x	0	Cannot classify

	2x	22	Brown (0.862)
	5x	19	Brown (0.484)
	10x	12	Brown (0.625)
36	MP	1	Not Brown (0.912)
	2x	0	Cannot classify
	5x	0	Cannot classify
	10x	0	Cannot classify
37	MP	1	Brown (0.686)
	2x	17	Brown (0.977)
	5x	9	Brown (0.972)
	10x	0	Cannot classify
38	MP	0	Cannot classify
	2x	0	Cannot classify
	5x	0	Cannot classify
	10x	0	Cannot classify

Table 9. Inferred 'most probable hair colour' for 38 samples, using the enrichment method (with the HIrisPlex webtool) at 2x, 5x, and 10x read depth thresholds. P-value scores are out of a value of 1, only the highest p-value is reported. Maximum number of phenotype SNPs for enrichment panel is 23.'NA' denotes samples for which most probable hair colour could not be determined. D-blond = Dark Blond, D-Brown = Dark Brown. Remaining p-values and AUC loss values are given in Supplementary File S6.

Sample	Coverage	No. of SNPs	Inferred hair colour (p-value)	Shade (p-value)	Most Probable Hair Colour
1	2x	20	Blond (0.657)	Light (0.935)	Blond/D-Blond
	5x	20	Blond (0.657)	Light (0.935)	Blond/D-Blond
	10x	20	Blond (0.657)	Light (0.935)	Blond/D-Blond
2	2x	23	Brown (0.526)	Light (0.842)	Brown/D-Brown
	5x	23	Brown (0.526)	Light (0.842)	Brown/D-Brown
	10x	23	Brown (0.526)	Light (0.842)	Brown/D-Brown
3	2x	20	Brown (0.533)	Light (0.863)	Brown/D-Brown
	5x	20	Brown (0.533)	Light (0.863)	Brown/D-Brown
	10x	20	Brown (0.533)	Light (0.863)	Brown/D-Brown
4	2x	0	Cannot classify	Cannot classify	NA
	5x	0	Cannot classify	Cannot classify	NA
	10x	0	Cannot classify	Cannot classify	NA
5	2x	0	Cannot classify	Cannot classify	NA
	5x	0	Cannot classify	Cannot classify	NA
	10x	0	Cannot classify	Cannot classify	NA
6	2x	23	Brown (0.443)	Light (0.896)	Brown/D-Brown
	5x	23	Brown (0.443)	Light (0.896)	Brown/D-Brown
	10x	23	Brown (0.443)	Light (0.896)	Brown/D-Brown
7	2x	23	Brown (0.561)	Light (0.834)	Brown/D-Brown
	5x	23	Brown (0.561)	Light (0.834)	Brown/D-Brown
	10x	23	Brown (0.561)	Light (0.834)	Brown/D-Brown
8	2x	21	Black (0.767)	Dark (0.989)	Black
	5x	10	Cannot classify	Cannot classify	NA
	10x	2	Cannot classify	Cannot classify	NA
9	2x	21	Black (0.769)	Dark (0.914)	Black
	5x	16	Cannot classify	Cannot classify	NA
	10x	7	Cannot classify	Cannot classify	NA
10	2x	23	Red (0.788)	Light (1)	Red
	5x	23	Red (0.788)	Light (1)	Red
	10x	19	Red (0.7671)	Light (1)	Red
11	2x	5	Cannot classify	Cannot classify	NA
	5x	2	Cannot classify	Cannot classify	NA
	10x	0	Cannot classify	Cannot classify	NA
12	2x	23	Brown (0.786)	Dark (0.921)	Brown/D-Brown
	5x	23	Brown (0.786)	Dark (0.921)	Brown/D-Brown

Sample	Coverage	No. of SNPs	Inferred eye colour (p-value)	Shade (p-value)	Most Probable Hair Colour
20	2x	3	Cannot classify	Cannot classify	NA
	5x	0	Cannot classify	Cannot classify	NA
	10x	0	Cannot classify	Cannot classify	NA
21	2x	20	Brown (0.595)	Light (0.806)	Brown/D-Brown
	5x	20	Brown (0.595)	Light (0.806)	Brown/D-Brown
	10x	18	Brown (0.578)	Light (0.796)	D-Brown/Black
22	2x	14	Red (0.547)	Light (0.76)	Red
	5x	11	Cannot classify	Cannot classify	NA
	10x	1	Cannot classify	Cannot classify	NA
23	2x	23	Black (0.582)	Dark (0.944)	Black
	5x	20	Black (0.58)	Dark (0.994)	Black
	10x	2	Cannot classify	Cannot classify	NA
24	2x	8	Cannot classify	Cannot classify	NA
	5x	1	Cannot classify	Cannot classify	NA
	10x	0	Cannot classify	Cannot classify	NA
25	2x	23	Blond (0.536)	Light (0.957)	Blond/D-Blond
	5x	23	Blond (0.536)	Light (0.957)	Blond/D-Blond
	10x	23	Blond (0.536)	Light (0.957)	Blond/D-Blond
26	2x	8	Cannot classify	Cannot classify	NA
	5x	2	Cannot classify	Cannot classify	NA
	10x	0	Cannot classify	Cannot classify	NA
27	2x	23	Blond (0.744)	Light (0.989)	Blond
	5x	23	Blond (0.744)	Light (0.989)	Blond
	10x	23	Blond (0.744)	Light (0.989)	Blond
28	2x	23	Blond (0.486)	Light (0.871)	D-Blond/Brown
	5x	23	Blond (0.486)	Light (0.871)	D-Blond/Brown
	10x	22	Blond (0.503)	Light (0.885)	D-Blond/Brown
29	2x	21	Blond (0.546)	Light (0.919)	Blond/D-Blond
	5x	18	Blond (0.601)	Light (0.933)	Blond/D-Blond
	10x	9	Cannot classify	Cannot Classify	NA
30	2x	23	Brown (0.473)	Light (0.936)	Brown/D-Brown
	5x	23	Brown (0.473)	Light (0.936)	Brown/D-Brown
	10x	23	Brown (0.473)	Light (0.936)	Brown/D-Brown
31	2x	0	Cannot classify	Cannot classify	NA
	5x	0	Cannot classify	Cannot classify	NA

	10x	23	Brown (0.786)	Dark (0.921)	Brown/D-Brown
13	2x	16	Brown (0.5)	Light (0.824)	Brown/D-Brown
	5x	7	Cannot classify	Cannot classify	NA
	10x	0	Cannot classify	Cannot classify	NA
14	2x	23	Blond (0.505)	Light (0.965)	Blond/D-Blond
	5x	23	Blond (0.505)	Light (0.965)	Blond/D-Blond
	10x	14	Blond (0.604)	Light (0.933)	Blond/D-Blond
15	2x	17	Brown (0.521)	Light (0.927)	Brown/D-Brown
	5x	9	Cannot classify	Cannot classify	NA
	10x	0	Cannot classify	Cannot classify	NA
16	2x	12	Cannot classify	Cannot classify	NA
	5x	2	Cannot classify	Cannot classify	NA
	10x	2	Cannot classify	Cannot classify	NA
17	2x	22	Brown (0.639)	Dark (0.569)	D-Brown/Black
	5x	20	Brown (0.667)	Dark (0.546)	D-Brown/Black
	10x	12	Brown (0.635)	Dark (0.701)	D-Brown/Black
18	2x	12	Cannot classify	Cannot classify	NA
	5x	4	Cannot classify	Cannot classify	NA
	10x	0	Cannot classify	Cannot classify	NA
19	2x	23	Blond (0.664)	Light (0.992)	Blond/D-Blond
	5x	23	Blond (0.664)	Light (0.992)	Blond/D-Blond
	10x	23	Blond (0.664)	Light (0.992)	Blond/D-Blond
	10x	0	Cannot classify	Cannot classify	NA
32	2x	0	Cannot classify	Cannot classify	NA
	5x	0	Cannot classify	Cannot classify	NA
	10x	0	Cannot classify	Cannot classify	NA
33	2x	18	Black (0.798)	Dark (0.996)	Black
	5x	5	Cannot classify	Cannot classify	NA
	10x	0	Cannot classify	Cannot classify	NA
34	2x	22	Brown (0.842)	Light (0.601)	Brown/D-Brown
	5x	21	Brown (0.821)	Light (0.541)	Brown/D-Brown
	10x	18	Brown (0.839)	Light (0.511)	Brown/D-Brown
35	2x	22	Brown (0.761)	Light (0.562)	Brown/D-Brown
	5x	19	Brown (0.773)	Dark (0.545)	Brown/D-Brown
	10x	12	Brown (0.719)	Light (0.516)	Brown/D-Brown
36	2x	0	Cannot classify	Cannot classify	NA
	5x	0	Cannot classify	Cannot classify	NA
	10x	0	Cannot classify	Cannot classify	NA
37	2x	17	Brown (0.57)	Dark (0.708)	D-Brown/Black
	5x	9	Brown (0.544)	Dark (0.788)	D-Brown/Black
	10x	0	Cannot classify	Cannot Classify	NA
38	2x	0	Cannot classify	Cannot classify	NA
	5x	0	Cannot classify	Cannot classify	NA
	10x	0	Cannot classify	Cannot classify	NA

4. Discussion

Environmentally challenged and degraded samples with low amounts of DNA can often fail to produce informative results with routine forensic genetic profiling due to high levels of damage and fragmentation. This study demonstrates the application of two recently developed SNP typing methods to various degraded and casework tooth and bone samples with extended post-mortem intervals in soil environments. The Miniplex recovers up to 18 SNPs to enable broad inferences for ancestry, eye colour, mtDNA, sex and Y-chr lineage. Subsequently, our 124-SNP custom hybridisation enrichment panel allows fine-resolution inferences for ancestry, hair and eye colour, sex and Y-chr lineage for degraded and casework samples. Forensic intelligence information and evidence gained using these methods are useful both in the selection and prioritisation of probative samples for downstream analysis, and for use in forensic investigation of degraded remains where no other biological information can be gathered.

Quantitative PCR (qPCR) assays are routine in forensic practice to determine the amount and suitability of a sample for STR typing (Lee *et al.* 2014). However, qPCR is an expensive, laborious and time-consuming process that is limited in comparatively assessing amplicons of varying sizes across mtDNA and nuclear targets in a simple assay. There is also no presumptive intelligence information that can be gained from these tests in order to screen samples based on biological profile. Hence, we used a combination of the Miniplex and Qubit quantification as an alternative to qPCR to determine sample quality and quantity prior to the application of hybridisation enrichment. The mitochondrial SNPs outperformed the nuclear markers, with full mitochondrial genotypes retrieved from all but one sample, compared with full nuclear SNP profiles from only a third of samples. This comparative assessment indicates advanced DNA degradation with limited availability of short nuclear DNA fragments, suggesting specialised methods such as hybridisation enrichment would be beneficial to recover nuclear SNPs. As the short DNA fragments (<128bp) targeted in the Miniplex were unable to be obtained from some samples, it is likely that current PCR-based MPS panels with SNP amplicons upwards of 200 bp would have limited success (Gettings *et al.* 2015; Churchill *et al.* 2016; Xavier & Parson 2017). The performance of the enrichment method to retrieve genetic information ranged from poor to obtaining full profiles and did not appear to be associated with DNA input. However, the performance of the enrichment panel generally aligned with SNP recovery using the Miniplex. That is, samples with no profile or partial Miniplex profiles retrieved either no, partial or full profiles using the enrichment

method, whereas those with high Miniplex success produced full profiles using the enrichment method. This demonstrates the value of the Miniplex in measuring DNA quality and predicting hybridisation enrichment success.

Sufficient SNPs could not be retrieved from some samples to make inferences of ancestry, phenotype, or paternal lineage (and sex) using the enrichment method. The low number of unique reads and high levels of clonality seen for these samples, suggest that the availability of targets was exhausted during the laboratory workflow. Future optimisation of techniques to maximise endogenous DNA input either at the DNA extraction, library preparation or enrichment stage is recommended to improve the success of the hybridisation enrichment-MPS strategy. The complexity and concentration of endogenous DNA used for library preparation appears to be the most critical factor influencing the success of genotyping by MPS techniques (Head *et al.* 2014; Sandoval-Velasco *et al.* 2017). The length of probe, variations of probe tiling and using RNA probes may also be viable adjustments to the hybridisation enrichment technique to explore (Cruz-Dávalos *et al.* 2017). However, compromised biological samples can contain short DNA fragments and a limited amount of starting material for genetic analysis (Dabney *et al.* 2013). In extreme cases, samples may only have the equivalent of a few intact diploid cells and any additional laboratory processing such as extraction, purification, library preparation and target enrichment could further reduce suitable DNA available for analysis due to inherent inefficiencies (van Oorschot *et al.* 2003; Aigrain *et al.* 2016; Chung *et al.* 2016; Cruz-Dávalos *et al.* 2017). The results of one hair shaft sample (sample 38) which did not retrieve any SNPs using the Miniplex (mtDNA or nuclear), and only three SNPs using hybridisation enrichment, would suggest that this sample had minimal amounts of endogenous DNA. Optimisation of current methods or future technical development may not improve the recovery of, nor increase the depth of coverage of targets if the amount of DNA available in a sample is extremely low. This may be the case for some of the samples in this study that recovered no, or very few SNPs at a low read depth.

Currently, there is no consensus with regards to read depth thresholds for SNP calling using MPS strategies amongst forensic practitioners, especially for hybridisation enrichment technologies where the removal of PCR duplicates is standard (Carpenter *et al.* 2013; Templeton *et al.* 2013; Samorodnitsky *et al.* 2015; García-García *et al.* 2016). The removal of duplicates, which can artificially inflate depth of coverage, prior to variant calling minimises the effect of PCR bias during library preparation, and reduces the risk of false positive calls (DePristo *et al.* 2011; Ebbert *et al.* 2016). In this study we measured SNP

typing success and resulting biological inferences based on three read depth thresholds after PCR duplicate removal. Low read depth coverage is a concern for genotype calling using MPS sequencing technologies given the risk of sequencing error and allele dropout. Although a higher read depth threshold is desired for the most confident SNP calling, in our study using a 10x threshold substantially reduced the number of SNPs from which interpretations could be made. Consequently, differing predictions for some samples were obtained at the 10x level with low statistical confidence compared to lower read depth thresholds, due to allele dropout with a higher read depth threshold. Future studies that focus on the application of various MPS typing methods on highly degraded casework samples are needed to reach a solid agreement on reporting criteria. It is important to emphasise the need for a standard set of guidelines that forensic investigators can use to report SNP data using MPS strategies (both PCR-based and alternative technologies). This is especially of concern for highly degraded casework samples where average read depth may be relatively low when removing PCR duplicates due to limited endogenous DNA content.

In two instances, discordance was noted between the genotypes generated by the two different methods (one Y-chr at 5x in one sample, and one autosomal biogeographic SNP at 80x read depth using MPS in another). A virtually even number of reads were obtained for the two alleles in the heterozygous biogeographic ancestry SNP using the hybridisation enrichment approach which could suggest dropout of one allele in the SNaPshot reaction. However, this did not result in conflicting biogeographic ancestry predictions for this sample. The discordant Y-chr SNP is more difficult to resolve given the sample showed high locus dropout and low read depth using both the Miniplex and enrichment method. The discrepancy could be due to SNaPshot mistyping or noise due to sub-optimal DNA input, or a result of sequencing error or incorrect read mapping using the enrichment method. The observation of discordant genotypes between SNaPshot and MPS has been demonstrated in previous studies, where it has been suggested that Sanger sequencing, or singleplex SNaPshot SNP typing of such loci can provide another avenue to assess genotype discrepancies (Daniel *et al.* 2015).

The typing of Y-chr SNPs using both methods in this study was useful for resolving paternal lineages (and ancestry), and for sex determination. A total of 36 Y-chr markers were typed across the two methods, and in no instance was sex predicted incorrectly for known female or male samples using either method. The only instances where sex could not be confidently predicted were for samples which lacked sufficient SNP data. Overall the results indicate that

the two methods are able to correctly identify sex in degraded samples. Owing to the larger number of Y-chr SNPs in the enrichment method, samples were able to be resolved into more specific Y-chr haplogroups than the Miniplex, allowing for finer resolution geographical affiliation of paternal lineages (i.e. Miniplex haplogroup 'Not D, E, C, R, O' into haplogroup G). For all samples where a Y-chr haplogroup was allocated using both methods, all assignments were phylogenetically concordant, indicating the methods are successfully able to infer paternal lineage in degraded samples for intelligence testing.

The assignment of biogeographic ancestry using the Miniplex and the enrichment method was generally concordant. However, five samples had conflicting predictions. Four of these samples generated a limited number of SNPs and very low and unreliable likelihood ratios using the Miniplex compared to the enrichment method. This result is not only influenced by the small number of SNPs available from the Miniplex for analysis, but also by the shared demographic history between some populations shown in previous studies where some alleles can exist in similar frequencies across both populations (Galanter *et al.* 2012; de la Puente *et al.* 2016). The final sample obtained a very small number of SNPs using both methods and had very weak likelihood ratios for both predictions and therefore could not be confidently assigned ancestry using any classification system. The combined use of Snipper, STRUCTURE and PCA in parallel has been advocated by ancestry analysis studies (de la Puente *et al*. 2016; Phillips *et al*. 2009) previously. Not unexpectedly, using the enrichment method generated a much higher number of SNPs and, therefore, higher likelihoods of ancestry assignment than the Miniplex. As the Miniplex was designed as a screening tool for broad ancestry predictions, inferences made using this panel should not be considered as final and should be followed by confirmatory ancestry testing (in this case, the enrichment method) for reporting of results. Overall, this study demonstrated the value of the two methods for inferring biogeographic ancestry in compromised samples to aid in forensic investigations.

A number of samples were able to return an eye or hair colour prediction using both methods, however problems arose for samples that, despite having an almost full profile, were missing SNPs that are the most informative for particular traits. For example, samples missing the rs12913832 SNP were unable to be assigned to an eye colour class using either method, even for samples retrieving up to 21 out of the 23 phenotype markers. The rs12913832 exists as one of the strongest predictors of eye colour (Sulem *et al.* 2008; Walsh *et al.* 2013), and the current IrisPlex and HIrisPlex model for inferring eye colour does not allow a classification if

this SNP is absent from the genotype. Considering that many of the phenotype markers are linked and existing on the same pigmentation genes, the single-tiled design of our MPS method results in a single probe to target multiple HIrisPlex SNPs (e.g. one probe targeting six phenotype SNPs on chromosome 16) (Chapter 3). A substantial loss in data could occur if this single probe in unsuccessful and will be detrimental to the prediction model to infer hair and eye colour. As a recommendation, it may be beneficial to use a tiled approach demonstrated in previous studies to increase the likelihood of enriching for multiple loci covered by a single probe (Cruz-Dávalos *et al.* 2017). Despite retrieving all targeted phenotype SNPs, eye colour predictions for two samples returned a statistical value lower than the suggested reporting threshold of 0.7 (Walsh *et al.* 2012; Walsh *et al.* 2014). This suggests a limitation with the current HIrisPlex SNP panel and prediction model already established rather than a limitation of the enrichment method used in this study. So far, only samples from European populations have been used to develop the prediction model (Walsh *et al.* 2014). Including samples of non-European origin will be beneficial in additional studies using the HIrisPlex system for model refinement. Further developments in DNA phenotyping and the identification of new pigmentation variants can also be included in the customisable enrichment panel in the future to improve the prediction of hair and eye colour.

Discussions on the use of ancestry testing for forensic intelligence have advocated for a more holistic genetic approach for determining biogeographic ancestry (Lao *et al.* 2010; Phillips 2015). The concerns around using a single biological query alone for determining ancestry has been demonstrated in previous studies (Phillips *et al.* 2009; Freire-Aradas *et al.* 2014), and combining uni-parental markers (Y-chr and mtDNA) with autosomal markers can better resolve ancestry classifications and admixture (Lao *et al.* 2010). The techniques in this study addressed this issue by combining multiple marker types for forensic ancestry information. In the present study, a sample indicating autosomal American ancestry with admixture from Europe was shown to carry a European paternal lineage using the enrichment method. The Miniplex inferred a broad mtDNA haplogroup spanning Eastern Eurasia, Southern Asia, Native America and Oceania (Bandelt *et al.* 2003; Quintana-Murci *et al.* 2004; Merriwether *et al.* 2005; Marrero *et al.* 2016), further confirming the contribution of multiple ancestral gene pools in this individual. Regardless of whether an individual has admixed ancestry, the genotyping of multiple marker types in both the Miniplex and enrichment panel adds extra components of genetic information through which a confident ancestry prediction can be made. The inclusion of mtDNA (in the Miniplex) and Y-chr SNPs in both the Miniplex and enrichment panel provides one (for females), or two (for males) possible independent sources

of ancestry inference which can help overcome the limitations of using autosomal SNPs alone. Combing multiple marker types in single panels also reduces the costs and labour demands of performing multiple independent tests each focusing on a single marker type and can also aid in minimising contamination risks during the workflow.

Overall, the results of this research suggest that the combined approach of the Miniplex and hybridisation capture panel holds potential to triage degraded samples for successful SNP typing, and to provide informative analysis of SNPs for intelligence information from degraded skeletal material.

5. Conclusion

In this study we applied two recently developed SNP typing approaches for estimating sample degradation and broad biological profile, and to retrieve forensically relevant intelligence SNPs for more specific inferences of biogeographic ancestry, paternal lineage, sex, and hair and eye colour from degraded and casework samples. We show the utility of this approach for obtaining intelligence data to aid in forensic investigations of degraded, historical and cold case samples including human hair and skeletal remains. However, the developing field of massively parallel sequencing for forensic purposes requires standard guidelines for the analysis and interpretation of sequencing data to maintain uniformity amongst forensic laboratories, especially for samples that are degraded or compromised.

Competing Interests

The authors declare no competing interests.

Acknowledgements

The authors acknowledge the South Australian Police Museum for access to hair samples. The research was supported by an Australian Research Council (ARC) Future Fellowship (FT10010008), ARC Discovery Project (DP150101664) and ARC LIEF Project (LE160100154) to JJA.

6. References

Adelaide News 1949, 'To Make Cast of Beach Body', <https://trove.nla.gov.au/newspaper/article/130194922>

Aigrain, L., Gu, Y. & Quail, M.A. 2016. Quantitation of next generation sequencing library preparation protocol efficiencies using droplet digital PCR assays - a systematic comparison of DNA library preparation kits for Illumina sequencing, *BMC genomics*, 17, 458.

Al-Asfi, M., McNevin, D., Mehta, B., Power, D., Gahan, M.E. & Daniel, R. 2018. Assessment of the Precision ID Ancestry panel, *Int J Legal Med*.

Bandelt, H.J., Herrnstadt, C., Yao, Y.G., Kong, Q.P., Kivisild, T., Rengo, C., Scozzari, R., Richards, M., Villems, R., Macaulay, V., et al. 2003. Identification of Native American Founder mtDNAs Through the Analysis of Complete mtDNA Sequences: Some Caveats, *Ann Hum Genet*, 67, 512-24.

Bardan, F., Higgins, D. & Austin, J.J. 2018. A mini-multiplex SNaPshot assay for the triage of degraded human DNA, *Forensic Sci Int Genet*, 34, 62-70.

Bose, N., Carlberg, K., Sensabaugh, G., Erlich, H. & Calloway, C. 2018. Target capture enrichment of nuclear SNP markers for massively parallel sequencing of degraded and mixed samples, *Forensic Sci Int Genet*, 34, 186-96.

Brotherton, P., Haak, W., Templeton, J., Brandt, G., Soubrier, J., Jane Adler, C., Richards, S.M., Sarkissian, C.D., Ganslmeier, R., Friederich, S., et al. 2013. Neolithic mitochondrial haplogroup H genomes and the genetic origins of Europeans, *Nat Commun*, 4, 1764.

Carpenter, Meredith L., Buenrostro, Jason D., Valdiosera, C., Schroeder, H., Allentoft, Morten E., Sikora, M., Rasmussen, M., Gravel, S., Guillén, S., Nekhrizov, G., et al. 2013. Pulling out the 1%: Whole-Genome Capture for the Targeted Enrichment of Ancient DNA Sequencing Libraries, *Am J Hum Genet*, 93, 852-64.

Chung, J., Son, D.-S., Jeon, H.-J., Kim, K.-M., Park, G., Ryu, G.H., Park, W.-Y. & Park, D. 2016. The minimal amount of starting DNA for Agilent's hybrid capture-based targeted massively parallel sequencing, *Sci Rep*, 6, 26732.

Churchill, J.D., Schmedes, S.E., King, J.L. & Budowle, B. 2016. Evaluation of the Illumina® Beta Version ForenSeq™ DNA Signature Prep Kit for use in genetic profiling, *Forensic Sci Int Genet*, 20, 20-9.

Cornelis, S., Fauvart, M., Gansemans, Y., Vander Plaetsen, A.-S., Colle, F., Wiederkehr, R.S., Deforce, D., Stakenborg, T. & Van Nieuwerburgh, F. 2018. Multiplex STR amplification sensitivity in a silicon microchip, *Sci Rep*, 8, 9853.

Cruz-Dávalos, D.I., Llamas, B., Gaunitz, C., Fages, A., Gamba, C., Soubrier, J., Librado, P., Seguin-Orlando, A., Pruvost, M., Alfarhan, A.H., et al. 2017. Experimental conditions improving in-solution target enrichment for ancient DNA, *Mol Ecol Resour*, 17, 508-22.

Dabney, J., Meyer, M. & Paabo, S. 2013. Ancient DNA damage, *Cold Spring Harb Perspect Biol*, 5.

Daniel, R., Santos, C., Phillips, C., Fondevila, M., van Oorschot, R.A.H., Carracedo, Á., Lareu, M.V. & McNevin, D. 2015. A SNaPshot of next generation sequencing for forensic SNP analysis, *Forensic Sci Int Genet*, 14, 50-60.

de la Puente, M., Phillips, C., Santos, C., Fondevila, M., Carracedo, Á. & Lareu, M.V. 2017. Evaluation of the Qiagen 140-SNP forensic identification multiplex for massively parallel sequencing, *Forensic Sci Int Genet*, 28, 35-43.

de la Puente, M., Santos, C., Fondevila, M., Manzo, L., Carracedo, A., Lareu, M.V. & Phillips, C. 2016. The Global AIMs Nano set: A 31-plex SNaPshot assay of ancestry-informative SNPs, *Forensic Sci Int Genet*, 22, 81-8.

DePristo, M.A., Banks, E., Poplin, R., Garimella, K.V., Maguire, J.R., Hartl, C., Philippakis, A.A., del Angel, G., Rivas, M.A., Hanna, M., et al. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data, *Nature Genet*, 43, 491.

Ebbert, M.T.W., Wadsworth, M.E., Staley, L.A., Hoyt, K.L., Pickett, B., Miller, J., Duce, J., for the Alzheimer's Disease Neuroimaging, I., Kauwe, J.S.K. & Ridge, P.G. 2016. Evaluating the necessity of PCR duplicate removal from next-generation sequencing data and a comparison of approaches, *BMC Bioinformatics*, 17, 239.

Edson, J., Brooks, E.M., McLaren, C., Robertson, J., McNevin, D., Cooper, A. & Austin, J.J. 2013. A quantitative assessment of a reliable screening technique for the STR analysis of telogen hair roots, *Forensic Sci Int Genet*, 7, 180-8.

Fondevila, M., Phillips, C., Santos, C., Freire Aradas, A., Vallone, P.M., Butler, J.M., Lareu, M.V. & Carracedo, Á. 2013. Revision of the SNPforID 34-plex forensic ancestry test: Assay enhancements, standard reference sample genotypes and extended population studies, *Forensic Sci Int Genet*, 7, 63-74.

Freire-Aradas, A., Ruiz, Y., Phillips, C., Maronas, O., Sochtig, J., Tato, A.G., Dios, J.A., de Cal, M.C., Silbiger, V.N., Luchessi, A.D., et al. 2014. Exploring iris colour prediction and ancestry inference in admixed populations of South America, *Forensic Sci Int Genet*, 13, 3-9.

Galanter, J.M., Fernandez-Lopez, J.C., Gignoux, C.R., Barnholtz-Sloan, J., Fernandez-Rozadilla, C., Via, M., Hidalgo-Miranda, A., Contreras, A.V., Figueroa, L.U., Raska, P., et al. 2012. Development of a Panel of Genome-Wide Ancestry Informative Markers to Study Admixture Throughout the Americas, *PLoS Genet*, 8, e1002554.

García-García, G., Baux, D., Faugère, V., Moclyn, M., Koenig, M., Claustres, M. & Roux, A.-F. 2016. Assessment of the latest NGS enrichment capture methods in clinical context, *Sci Rep*, 6, 20948.

Gettings, K.B., Kiesler, K.M. & Vallone, P.M. 2015. Performance of a next generation sequencing SNP assay on degraded DNA, *Forensic Sci Int Genet*, 19, 1-9.

Head, S.R., Komori, H.K., LaMere, S.A., Whisenant, T., Van Nieuwerburgh, F., Salomon, D.R. & Ordoukhanian, P. 2014. Library construction for next-generation sequencing: overviews and challenges, *Biotechniques*, 56, 61-passim.

Higgins, D., Rohrlach, A.B., Kaidonis, J., Townsend, G. & Austin, J.J. 2015. Differential Nuclear and Mitochondrial DNA Preservation in Post-Mortem Teeth with Implications for Forensic and Ancient DNA Studies, *PLoS One*, 10, e0126935.

Kline, M.C., Duewer, D.L., Redman, J.W. & Butler, J.M. 2005. Results from the NIST 2004 DNA quantitation study, *J Forensic Sci*, 50, 571-8.

Lao, O., Vallone, P.M., Coble, M.D., Diegoli, T.M., van Oven, M., van der Gaag, K.J., Pijpe, J., de Knijff, P. & Kayser, M. 2010. Evaluating Self-declared Ancestry of U.S. Americans with Autosomal, Y-chromosomal and Mitochondrial DNA, *Human Mutat*, 31, e1875-e93.

Lee, S.B., McCord, B. & Buel, E. 2014. Advances in forensic DNA quantification: A review, *Electrophoresis*, 35, 3044-52.

Marrero, P., Abu-Amero, K.K., Larruga, J.M. & Cabrera, V.M. 2016. Carriers of human mitochondrial DNA macrohaplogroup M colonized India from southeastern Asia, *BMC Evolutionary Biology*, 16, 246.

Merriwether, D.A., Hodgson, J.A., Friedlaender, F.R., Allaby, R., Cerchio, S., Koki, G. & Friedlaender, J.S. 2005. Ancient mitochondrial M haplogroups identified in the Southwest Pacific, *Proceedings of the National Academy of Sciences of the United States of America*, 102, 13034.

Musgrave-Brown, E., Ballard, D., Balogh, K., Bender, K., Berger, B., Bogus, M., Børsting, C., Brion, M., Fondevila, M., Harrison, C., et al. 2007. Forensic validation of the SNPforID 52-plex assay, *Forensic Sci Int Genet*, 1, 186-90.

Phillips, C. 2015. Forensic genetic analysis of bio-geographical ancestry, *Forensic Sci Int Genet*, 18, 49-65.

Phillips, C., Prieto, L., Fondevila, M., Salas, A., Gómez-Tato, A., Álvarez-Dios, J., Alonso, A., Blanco-Verea, A., Brión, M., Montesino, M., et al. 2009. Ancestry Analysis in the 11-M Madrid Bomb Attack Investigation, *PLoS One*, 4, e6583.

Porras-Hurtado, L., Ruiz, Y., Santos, C., Phillips, C., Carracedo, A. & Lareu, M.V. 2013. An overview of STRUCTURE: applications, parameter settings, and supporting software, *Front Genet*, 4, 98.

Quintana-Murci, L., Chaix, R., Wells, R.S., Behar, D.M., Sayar, H., Scozzari, R., Rengo, C., Al-Zahery, N., Semino, O., Santachiara-Benerecetti, A.S., et al. 2004. Where West Meets East: The Complex mtDNA Landscape of the Southwest and Central Asian Corridor, *The American Journal of Human Genetics*, 74, 827-45.

Samorodnitsky, E., Datta, J., Jewell, B.M., Hagopian, R., Miya, J., Wing, M.R., Damodaran, S., Lippus, J.M., Reeser, J.W., Bhatt, D., et al. 2015. Comparison of Custom Capture for Targeted Next-Generation DNA Sequencing, *The Journal of Molecular Diagnostics : JMD*, 17, 64-75.

Sandoval-Velasco, M., Lundstrøm, I.K.C., Wales, N., Ávila-Arcos, M.C., Schroeder, H. & Gilbert, M.T.P. 2017. Relative performance of two DNA extraction and library preparation methods on archaeological human teeth samples, *STAR: Science & Technology of Archaeological Research*, 3, 80-8.

Sulem, P., Gudbjartsson, D.F., Stacey, S.N., Helgason, A., Rafnar, T., Jakobsdottir, M., Steinberg, S., Gudjonsson, S.A., Palsson, A., Thorleifsson, G., et al. 2008. Two newly identified genetic determinants of pigmentation in Europeans, *Nature Genet*, 40, 835.

Templeton, J.E.L., Brotherton, P.M., Llamas, B., Soubrier, J., Haak, W., Cooper, A. & Austin, J.J. 2013. DNA capture and next-generation sequencing can recover whole mitochondrial genomes from highly degraded samples for human identification, *Investig Genet*, 4, 26.

van Oorschot, R.A.H., Phelan, D.G., Furlong, S., Scarfo, G.M., Holding, N.L. & Cummins, M.J. 2003. Are you collecting all the available DNA from touched objects?, *International Congress Series*, 1239, 803-7.

Walsh, S., Chaitanya, L., Clarisse, L., Wirken, L., Draus-Barini, J., Kovatsi, L., Maeda, H., Ishikawa, T., Sijen, T., de Knijff, P., et al. 2014. Developmental validation of the HIrisPlex system: DNA-based eye and hair colour prediction for forensic and anthropological usage, *Forensic Sci Int Genet*, 9, 150-61.

Walsh, S., Liu, F., Wollstein, A., Kovatsi, L., Ralf, A., Kosiniak-Kamysz, A., Branicki, W. & Kayser, M. 2013. The HIrisPlex system for simultaneous prediction of hair and eye colour from DNA, *Forensic Sci Int Genet*, 7, 98-115.

Walsh, S., Wollstein, A., Liu, F., Chakravarthy, U., Rahu, M., Seland, J.H., Soubrane, G., Tomazzoli, L., Topouzis, F., Vingerling, J.R., et al. 2012. DNA-based eye colour prediction across Europe with the IrisPlex system, *Forensic Sci Int Genet*, 6, 330-40.

Xavier, C. & Parson, W. 2017. Evaluation of the Illumina ForenSeq™ DNA Signature Prep Kit – MPS forensic application for the MiSeq FGx™ benchtop sequencer, *Forensic Sci Int Genet*, 28, 188-94.

7. Supplementary Files

Supplementary File S1. Details of the 402 reference samples from five population groups across the 1000 Genomes and HGDP-CEPH datasets used for comparison to Miniplex data. All genotypes are provided in the strand direction of primers used in the Miniplex for direct comparison (provided as an electronic file on USB Drive).

Supplementary File S2. Details of the 368 reference samples from five population groups across the 1000 Genomes and HGDP-CEPH datasets used for comparison to MPS data. All genotypes are presented in forward direction (provided as an electronic file on USB Drive).

Supplementary Table S3. Final DNA input amount for Miniplex PCR and hybridisation enrichment reactions for 38 degraded teeth and casework samples. Input for Miniplex PCR was quantified using Qubit High Sensitivity assay on DNA extracts. Input for hybridisation enrichment reactions was quantified using Qubit High Sensitivity assay on DNA library preparation products.

Sample	Sex	Sample Type	Sample Description	DNA Input for Miniplex (ng)	DNA Input for Enrichments (ng)
1	Male	Degraded Teeth	Tooth	36.2	310.2
2	Female	Degraded Teeth	Tooth	0.723	322.4
3	Female	Degraded Teeth	Tooth	26.6	303.6
4	Male	Degraded Teeth	Tooth	0.128	6.2
5	Male	Degraded Teeth	Tooth	0.097	390
6	Male	Degraded Teeth	Tooth	0.336	212.7
7	Female	Degraded Teeth	Tooth	0.812	317.2
8	Female	Degraded Teeth	Tooth	0.13	507
9	Female	Degraded Teeth	Tooth	0.404	122.7
10	Female	Degraded Teeth	Tooth	0.297	439.4
11	Male	Degraded Teeth	Tooth	<0.5	145.6
12	Male	Degraded Teeth	Tooth	0.493	327.6
13	Female	Degraded Teeth	Tooth	1.14	429
14	Female	Degraded Teeth	Tooth	0.505	314.6
15	Female	Degraded Teeth	Tooth	0.401	83
16	Female	Degraded Teeth	Tooth	0.598	189.8
17	Female	Degraded Teeth	Tooth	0.283	260
18	Female	Degraded Teeth	Tooth	0.188	325
19	Male	Degraded Teeth	Tooth	0.552	341.9
20	Male	Degraded Teeth	Tooth	0.722	249.6
21	Male	Degraded Teeth	Tooth	0.638	122.84
22	Male	Degraded Teeth	Tooth	1.03	390
23	Male	Degraded Teeth	Tooth	<0.5	358.8
24	Male	Degraded Teeth	Tooth	<0.5	197.6
25	Female	Degraded Teeth	Tooth	0.507	444.6
26	Male	Degraded Teeth	Tooth	0.139	546
27	Female	Degraded Teeth	Tooth	0.36	553.8
28	Female	Degraded Teeth	Tooth	0.125	451.8
29	Female	Degraded Teeth	Tooth	0.167	403
30	Male	Degraded Teeth	Tooth	0.102	325
31	UNK	Casework	Bone	0.154	23.33
32	UNK	Casework	Bone	0.238	516
33	UNK	Casework	Bone	0.506	501
34	UNK	Casework	Bone	0.408	504
35	UNK	Casework	Bone	0.671	513.6
36	UNK	Casework	Bone	0.263	46.4
37	Male	Casework	Anagen hair	0.392	504
38	Male	Casework	Hair shaft	0.109	101.5

Supplementary Table S4. Total SNP recovery and Y-chr SNP recovery of Miniplex and custom enrichment panel on 38 degraded human teeth and casework samples with varying DNA input amounts. SNP typing success for custom enrichment panel is reported for a minimum read depth of coverage of 2x, 5x, and 10x.

Sample	Sex	DNA Input (ng)	Coverage	SNP Recovery (%)	Y-SNP Recovery (%)
1	M	36.2	Miniplex	100	100
			2x	95.2	100
		310.2	5x	95.2	100
			10x	95.2	100
2	F	0.723	Miniplex	84.6	
			2x	100	
		322.4	5x	100	
			10x	100	
3	F	26.6	Miniplex	100	
			2x	93.2	
		303.6	5x	93.2	
			10x	93.2	
4	M	0.128	Miniplex	44.4	20
			2x	1.6	2.9
		6.2	5x	0.8	2.9
			10x	0	2.9
5	M	0.097	Miniplex	44.4	20
			2x	3.2	0
		390	5x	0	0
			10x	0	0
6	M	0.336	Miniplex	66.6	40
			2x	100	100
		212.7	5x	96.8	88.6
			10x	83.1	45.7
7	F	0.812	Miniplex	100	
			2x	100	
		317.2	5x	100	
			10x	100	
8	F	0.13	Miniplex	92	
			2x	91	
		507	5x	47.2	
			10x	6.7	
14	F	0.505	Miniplex	100	
			2x	100	
		310.2	5x	95.5	
			10x	60.7	
15	F	0.401	Miniplex	53.8	
			2x	71.9	
		322.4	5x	22.5	
			10x	0	
16	F	0.598	Miniplex	69.2	
			2x	48.3	
		303.6	5x	12.4	
			10x	4.5	
17	F	0.283	Miniplex	84.6	
			2x	96.6	
		6.2	5x	88.8	
			10x	40.4	
18	F	0.188	Miniplex	69.2	
			2x	65.2	
		390	5x	15.7	
			10x	1.1	
19	M	0.552	Miniplex	100	100
			2x	100	100
		212.7	5x	100	100
			10x	100	100
20	M	0.722	Miniplex	61.1	80
			2x	17.7	14.3
		317.2	5x	6.5	8.6
			10x	1.6	0
21	M	0.638	Miniplex	100	100
			2x	93.5	94.2
		507	5x	82.3	57.1
			10x	47.6	8.6
27	F	0.36	Miniplex	100	
			2x	100	
		310.2	5x	100	
			10x	97.8	
28	F	0.125	Miniplex	100	
			2x	100	
		322.4	5x	100	
			10x	100	
29	F	0.167	Miniplex	61.5	
			2x	96.6	
		303.6	5x	79.8	
			10x	31.5	
30	M	0.102	Miniplex	94.4	100
			2x	100	100
		6.2	5x	98.4	97.1
			10x	98.4	97.1
31	UNK	0.154	Miniplex	33.3	20
			2x	0	0
		390	5x	0	0
			10x	0	0
32	UNK	0.238	Miniplex	61.1	60
			2x	13.7	2.9
		212.7	5x	4	0
			10x	0.8	0
33	UNK	0.506	Miniplex	55.5	60
			2x	38.7	17.1
		317.2	5x	11.3	0
			10x	0	0
34	UNK	0.408	Miniplex	100	100
			2x	94.4	91.4
		507	5x	83.1	65.7
			10x	55.6	25.7

9	F	0.404	Miniplex	69.2	
			2x	84.3	
		122.7	5x	50.6	
			10x	27	
10	F	0.297	Miniplex	84.6	
			2x	100	
		439.4	5x	100	
			10x	66.3	
11	M	<0.05	Miniplex	72.2	40
			2x	27.4	17.1
		145.6	5x	12.9	11.4
			10x	1.6	5.8
12	M	0.493	Miniplex	100	100
			2x	100	100
		327.6	5x	100	100
			10x	100	100
13	F	1.14	Miniplex	76.9	
			2x	71.9	
		429	5x	37.1	
			10x	7.9	

22	M	1.03	Miniplex	100	100
			2x	49.2	25.7
		122.7	5x	21.8	0
			10x	0.8	0
23	M	<0.05	Miniplex	33.3	0
			2x	95.2	82.9
		439.4	5x	71.8	31.4
			10x	18.5	2.9
24	M	<0.05	Miniplex	38.8	40
			2x	25.8	14.3
		145.6	5x	7.3	5.8
			10x	1.6	0
25	F	0.507	Miniplex	100	
			2x	100	
		327.6	5x	100	
			10x	100	
26	M	0.139	Miniplex	38.8	20
			2x	9.7	0
		429	5x	1.6	0
			10x	0	0

35	UNK	0.671	Miniplex	94.4	100
			2x	96	91.4
		122.7	5x	76.6	60
			10x	40.3	17.1
36	UNK	0.263	Miniplex	66.6	40
			2x	0	0
		439.4	5x	0	0
			10x	0	0
37	M	0.392	Miniplex	61.1	60
			2x	36.3	45.7
		145.6	5x	26.6	17.1
			10x	3.2	0
38	M	0.109	Miniplex	0	0
			2x	2.4	2.9
		327.6	5x	0.8	0
			10x	0	0

Supplementary File S5. Inferred predictions for 38 degraded teeth and casework samples using the Miniplex and enrichment panel at 2x, 5x, and 10x coverage thresholds. Maximum number of ancestry SNPs that can be obtained from the Miniplex is five. Maximum number of ancestry SNPs for enrichment panel is 67. Maximum number of phenotype SNPs that can be obtained from Miniplex is three. Maximum number of phenotype SNPs for enrichment panel is 23. Maximum number of Y-chr SNPs that can be obtained from Miniplex is five. Maximum number of Y-chr SNPs for enrichment panel is 35. 'NA' denotes samples were no prediction could be made due to a lack of SNPs. (provided as an electronic copy on USB drive).

Supplementary File S6. Major ancestry component and average population membership proportions from STRUCTURE analysis of 38 degraded teeth and casework samples using the Miniplex and enrichment panel at 2x, 5x, and 10x read depth thresholds (K = 5). Maximum number of ancestry SNPs that can be obtained using the Miniplex is five. Maximum number of SNPs for enrichment panel is 67.

Sample	Coverage	SNPs	Major Ancestry	AFR %	AMR %	EAS %	EUR %	OCE %
1	Miniplex	5	EUR	1.23	0.67	0.93	96.50	0.70
	2x	64	EUR	0.73	0.23	0.20	91.87	6.93
	5x	64	EUR	0.63	0.30	0.20	91.97	6.93
	10x	64	EUR	0.53	0.27	0.13	92.07	6.93
2	Miniplex	3	EUR	1.23	0.87	1.07	95.93	0.83
	2x	67	EUR	1.70	1.03	4.07	89.27	3.97
	5x	67	EUR	2.07	22.07	0.20	74.87	0.80
	10x	67	AFR	40.43	18.07	2.30	23.10	16.07
3	Miniplex	5	EUR	1.13	0.63	0.87	93.43	3.93
	2x	64	EUR	1.90	1.10	3.33	90.37	3.30
	5x	64	EUR	1.67	1.23	4.03	89.97	3.13
	10x	64	EUR	1.87	1.20	3.67	90.27	3.00
4	Miniplex	0	NA	-	-	-	-	-
	2x	1	EUR	1.53	21.63	25.80	31.73	19.27
	5x	0	NA	-	-	-	-	-
	10x	0	NA	-	-	-	-	-
5	Miniplex	2	EUR	43.17	2.20	2.33	50.53	1.87
	2x	4	EUR	23.40	1.17	1.07	73.97	0.40
	5x	0	NA	-	-	-	-	-
	10x	0	NA	-	-	-	-	-
6	Miniplex	2	EUR	2.23	1.13	1.57	93.73	1.33
	2x	67	EUR	0.43	1.40	0.20	96.40	1.53
	5x	67	EUR	1.63	1.10	4.47	89.10	3.77
	10x	65	EUR	1.97	1.27	4.30	89.00	3.47
7	Miniplex	5	EUR	1.23	0.63	0.90	96.50	0.70
	2x	67	EUR	0.20	0.10	0.10	99.43	0.13
	5x	67	EUR	0.27	0.10	0.20	99.33	0.17
	10x	67	EUR	0.20	0.10	0.13	99.40	0.13
8	Miniplex	4	AMR	2.67	53.13	33.43	3.37	7.37
	2x	61	AMR	0.47	60.63	25.43	13.17	0.33
	5x	33	AMR	0.50	95.93	2.43	0.80	0.33
	10x	5	AMR	1.93	71.83	2.10	23.03	1.17
9	Miniplex	2	EAS	17.10	28.93	29.40	5.10	19.47
	2x	55	AMR	0.63	85.53	0.97	12.63	0.27
	5x	29	AMR	1.90	68.07	4.50	24.83	0.73
	10x	17	AMR	0.80	90.27	3.17	5.20	0.53
10	Miniplex	4	EUR	1.33	1.07	1.57	95.33	0.73
	2x	67	EUR	0.43	0.23	0.10	93.27	5.93
	5x	67	EUR	0.37	0.20	0.13	93.10	6.17
	10x	40	EUR	0.20	0.10	0.13	85.60	13.93
11	Miniplex	3	EUR	2.30	0.93	0.97	94.93	0.80
	2x	23	EUR	3.83	0.80	3.40	90.63	1.40
	5x	11	EUR	0.60	0.30	0.63	97.57	0.97
	10x	0	NA	-	-	-	-	-
12	Miniplex	5	EUR	1.20	0.70	0.90	96.47	0.77
	2x	67	EUR	0.10	0.10	0.10	99.57	0.10
	5x	67	EUR	0.13	0.10	0.10	99.57	0.10

	10x	67	EUR	0.10	0.10	0.10	99.60	0.10
13	Miniplex	2	EUR	2.67	2.00	2.63	60.23	32.43
	2x	48	EUR	0.20	0.23	10.83	87.73	1.00
	5x	28	EUR	0.20	0.57	16.57	82.03	0.60
	10x	7	EUR	0.57	2.67	0.60	94.50	1.67
14	Miniplex	5	EUR	1.30	0.70	0.90	96.43	0.67
	2x	67	EUR	0.10	0.20	1.80	97.67	0.23
	5x	64	EUR	0.30	0.30	1.00	98.17	0.20
	10x	40	EUR	0.27	0.27	0.87	98.40	0.13
15	Miniplex	2	EUR	2.30	1.80	2.00	93.17	0.80
	2x	47	EUR	0.10	0.60	9.80	89.23	0.27
	5x	13	EUR	0.37	0.43	10.90	87.30	1.03
	10x	0	NA	-	-	-	-	-
16	Miniplex	1	EUR	41.67	1.93	2.20	52.47	1.73
	2x	32	EUR	1.70	12.03	0.70	85.23	0.27
	5x	9	EUR	4.10	14.43	6.57	73.47	1.47
	10x	2	OCE	2.53	23.77	10.13	8.43	55.13
17	Miniplex	3	EUR	1.20	1.63	1.97	94.40	0.83
	2x	65	EUR	0.13	0.10	2.90	96.47	0.40
	5x	56	EUR	0.17	0.13	3.70	95.13	0.83
	10x	24	EUR	0.43	0.20	6.83	90.77	1.77
18	Miniplex	2	EUR	2.20	1.00	1.00	94.27	1.53
	2x	46	EUR	0.67	0.17	0.17	97.37	1.63
	5x	11	EUR	1.37	0.67	1.00	95.73	1.30
	10x	1	AFR	37.07	1.23	17.90	27.07	16.73
19	Miniplex	5	EUR	1.27	0.73	0.93	96.40	0.67
	2x	67	EUR	0.27	0.20	0.30	99.13	0.10
	5x	67	EUR	0.23	0.20	0.27	99.17	0.10
	10x	67	EUR	0.20	0.20	0.23	99.23	0.10
20	Miniplex	2	EUR	39.23	1.10	1.50	42.67	15.43
	2x	14	EUR	0.33	0.33	0.27	98.87	0.20
	5x	5	EUR	1.47	0.57	0.97	83.07	13.97
	10x	2	AFR	65.13	1.17	2.07	23.50	8.10
21	Miniplex	5	EUR	1.17	0.67	0.83	96.60	0.73
	2x	64	EUR	0.67	0.10	0.20	97.63	1.40
	5x	63	EUR	0.73	0.13	0.17	97.50	1.47
	10x	41	EUR	1.43	0.17	0.20	96.60	1.57
22	Miniplex	5	EUR	1.17	0.67	0.83	96.60	0.73
	2x	38	EUR	0.67	0.10	0.20	97.63	1.40
	5x	16	EUR	0.73	0.13	0.17	97.50	1.47
	10x	1	AFR	1.10	1.53	2.07	94.57	0.80
23	Miniplex	0	NA	-	-	-	-	-
	2x	67	AMR	0.20	62.13	0.37	35.97	1.33
	5x	58	AMR	0.23	57.57	0.53	38.27	3.40
	10x	20	AMR	2.13	75.57	2.43	17.23	2.60
24	Miniplex	0	NA	-	-	-	-	-
	2x	19	EUR	30.80	5.67	3.80	59.33	0.37
	5x	6	EAS	4.40	20.03	65.50	8.47	1.67
	10x	2	AMR	10.07	70.07	14.20	3.77	1.90
25	Miniplex	5	EUR	1.03	0.67	0.87	93.73	3.70
	2x	67	EUR	0.23	0.23	0.27	99.03	0.20
	5x	67	EUR	0.20	0.30	0.23	99.03	0.23
	10x	67	EUR	0.20	0.23	0.20	99.10	0.20
26	Miniplex	1	EUR	2.23	1.60	1.97	92.70	1.47
	2x	4	EUR	7.57	3.00	9.40	77.77	2.27
	5x	2	AFR	71.37	1.77	3.13	20.47	3.30
	10x	2	AFR	69.70	1.67	3.37	21.90	3.37
27	Miniplex	5	EUR	1.27	0.70	0.80	96.50	0.67
	2x	67	EUR	0.20	0.17	0.20	99.20	0.20
	5x	67	EUR	0.17	0.17	0.23	99.27	0.17
	10x	65	EUR	0.10	0.17	0.27	99.37	0.10
28	Miniplex	5	EUR	1.17	0.70	1.03	96.30	0.83

	2x	67	EUR	0.30	0.27	1.07	98.20	0.17
	5x	67	EUR	0.37	1.40	0.23	96.33	1.67
	10x	67	EUR	0.37	1.30	0.20	96.13	1.93
29	Miniplex	1	EUR	1.10	1.73	2.10	93.47	1.53
	2x	65	EUR	0.17	0.80	12.53	86.37	0.13
	5x	52	EUR	0.20	0.50	17.20	81.97	0.13
	10x	18	EUR	1.73	1.40	2.60	93.20	1.07
30	Miniplex	4	EUR	0.10	0.17	11.33	87.93	0.43
	2x	67	EUR	0.17	0.17	10.13	88.70	0.83
	5x	67	EUR	0.13	0.23	9.90	88.97	0.77
	10x	66	EUR	1.10	0.90	1.03	96.13	0.77
31	Miniplex	0	NA	-	-	-	-	-
	2x	0	NA	-	-	-	-	-
	5x	0	NA	-	-	-	-	-
	10x	0	NA	-	-	-	-	-
32	Miniplex	2	EUR	38.70	7.60	13.17	38.93	1.53
	2x	15	EUR	0.73	0.23	0.40	98.17	0.47
	5x	5	EUR	1.17	1.50	22.63	73.87	0.77
	10x	1	AFR	32.33	19.13	17.60	29.40	1.53
33	Miniplex	1	EUR	2.23	20.53	24.00	33.73	19.50
	2x	34	EAS	0.20	0.23	98.93	0.20	0.40
	5x	10	EAS	1.10	1.60	87.90	8.43	0.93
	10x	0	NA	-	-	-	-	-
34	Miniplex	5	EUR	1.30	0.70	0.97	96.33	0.70
	2x	64	EUR	2.13	0.60	0.80	96.07	0.47
	5x	60	EUR	2.10	0.73	1.03	95.60	0.53
	10x	43	EUR	3.67	0.67	1.33	93.53	0.83
35	Miniplex	5	EUR	1.17	0.70	0.87	96.60	0.67
	2x	66	EUR	0.20	0.10	0.10	99.53	0.10
	5x	55	EUR	0.23	0.10	0.13	99.43	0.10
	10x	28	EUR	0.30	0.20	0.23	99.03	0.20
36	Miniplex	3	EUR	1.17	0.67	0.87	96.00	1.27
	2x	0	NA	-	-	-	-	-
	5x	0	NA	-	-	-	-	-
	10x	0	NA	-	-	-	-	-
37	Miniplex	2	EUR	2.30	1.53	2.07	93.30	0.77
	2x	46	EUR	0.87	0.10	0.27	98.47	0.30
	5x	17	EUR	2.67	0.20	0.37	96.33	0.37
	10x	4	EUR	2.37	0.77	1.00	94.83	1.03
38	Miniplex	0	NA	-	-	-	-	-
	2x	2	OCE	41.23	1.97	1.67	1.73	53.40
	5x	1	NA	29.37	21.00	27.43	4.07	18.13
	10x	0	NA	-	-	-	-	-

Chapter 5

The Historical Australian DNA Database

Manuscript prepared for submission

Statement of Authorship

Title of Paper	The Historical Australian DNA Database
Publication Status	☐ Published ☐ Accepted for Publication ☐ Submitted for Publication ☑ Unpublished and Unsubmitted w ork w ritten in manuscript style
Publication Details	Manuscript prepared for publication

Principal Author

Name of Principal Author (Candidate)	Felicia Bardan
Contribution to the Paper	Helped conceive the study, helped collect the samples, helped generate, analyse and interpret the data, drafted the manuscript and produced the figures.
Overall percentage (%)	60%
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.
Signature	Date 19/10/18

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

i. the candidate's stated contribution to the publication is accurate (as detailed above);

ii. permission is granted for the candidate in include the publication in the thesis; and

iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author	Denice Higgins
Contribution to the Paper	Helped generate and analyse data, revised the manuscript
Signature	Date 20/10/18

Name of Co-Author	Jeremy J Austin
Contribution to the Paper	Helped conceive the study, helped collect the samples, helped generate and analyse and interpret the data, revised the manuscript and helped produced the figures.
Signature	Date 18/10/18

The Historical Australian DNA Database

Felicia Bardan[1], Denice Higgins[1], Jeremy J. Austin[1]

[1]Australian Centre for Ancient DNA, School of Biological Sciences, University of Adelaide, Adelaide, South Australia 5005, Australia

*corresponding author: felicia.bardan@adelaide.edu.au

Abstract

Efforts to repatriate the remains of deceased Australian military personnel from World War I and II are hindered by a lack of contemporaneous biogeographic ancestry data for the Australian population. Extensive post-mortem damage and loss of anthropologically-informative skeletal elements limits identification of remains using traditional forensic methods such as short tandem repeat (STR) analysis, anthropology and odontology. Therefore, alternative DNA analyses are required to predict the biogeographic ancestry of remains to distinguish Australian soldiers in the first instance. This is a crucial step towards the correct repatriation and subsequent positive identification of remains. However, the biogeographic ancestry composition of the Australian population during the early 20th Century is largely unknown, and this may impact on the accuracy and reliability of sorting of recovered remains based on genetic ancestry. This study details the construction of the Historical Australian DNA Database (HADD), designed to resolve the genetic ancestry components of the historical Australian wartime population using mtDNA control region data and autosomal ancestry SNPs. An initial set of 259 individuals was analysed and was found to have predominantly European ancestry, with the true proportion of non-European ancestry in the Australian population before 1945 likely to be up to 2.8%. Based on autosomal SNP data, no sample was predicted to be predominantly non-European, whilst for mtDNA two samples carried non-European haplogroups, namely Australian Aboriginal S1 and Native American founder lineage B2d. Our results demonstrate the value of a multi-faceted approach to biogeographic ancestry prediction and the power of combining lineage and autosomal data. This approach has produced new ancestry data and generates the foundations of the first multi-gene historical DNA database for Australia. This database allows for a greater understanding of genetic admixture in the Australian population during the World War eras and can be used to objectively evaluate results from ancestry identification of Australian historical remains.

Keywords: forensic database, historical Australian population, mtDNA, autosomal ancestry informative makers, SNP

Introduction

Australia's involvement across both World War I (WWI, 1914-18) and World War II (WWII, 1939-1945) saw almost 1.5 million men and women serving. Australian forces participated in numerous campaigns across Europe, Northern Africa and close to Australian shores against Japanese forces in the Asia-Pacific region. Although experiencing many victories, battles were costly with over 100 thousand casualties in the Australian War Memorial Roll of Honour (Australian War Memorial 2017a). Such was the nature of the wars that large numbers of casualties were buried in unmarked graves, as unidentified servicemen in Commonwealth war graves or were never recovered due to difficult terrain or dangerous battle environments. Consequently, a significant number of servicemen are still reported as missing (assumed deceased). Although there is no official figure, approximately 35,000 Australian servicemen from WWI and WWII are estimated to rest in unmarked graves or unknown burial sites (Department of Veterans Affairs 2016). The Australian Defence Force (ADF) continues to investigate the discovery of human remains across historic battlefields in an effort to repatriate and identify fallen Australian servicemen from various conflicts.

Identification of remains is the ultimate goal of all war dead recovery efforts. However, positive identification of remains is a difficult and complex process that is hampered by the time elapsed since death (>70 years), and difficulty in locating ante-mortem information and family relatives for comparison. Furthermore, the long post-mortem interval and sub-optimal preservation conditions lead to severe degradation of biological remains, often combined with scattering, co-mingling and predation. Typically, the first step in the identification process relies on determining whether or not remains may belong to an Australian soldier. This process can be assisted by estimating the geographic origin or biogeographic ancestry. Remains believed to be of an allied or enemy soldier are respectfully repatriated to relevant governments, while those thought to belong to Australian soldiers are retained for further identity testing. Accurate determination of biogeographic ancestry is a critical part of the repatriation process to ensure remains are correctly returned to their country of origin.

Current forensic methods of identification are often limited in their ability to classify or group historical remains into a country of origin. Where possible, conventional methods of identification based on anthropology and dental records are used but rely on the presence of diagnostic body features such as the skull and teeth which are often not available. For ancestry determination, variation between ethnic/ancestral groups manifests as morphological differences in cranial features (Church 1995) and dental characteristics (Yaacob *et al.* 1996;

Edgar 2013). Personal effects and artefacts such as dog tags, uniforms and weapons can also assist in identification, or at least inform from which of the armed forces they may belong to. However, close quarters combat can make distinguishing between individuals extremely difficult due to widespread co-mingling. Additionally, remains can be fragmentary and dispersed due to removal from their original resting place by locals and/or predators. In these circumstances, such methods are of limited value, resulting in a reliance upon DNA-based techniques.

Ancestral population groups exhibit genetic variation between each other, allowing distinction based on geographic origin (Phillips 2015; de la Puente *et al.* 2016). DNA-based methods for ancestry determination include mitochondrial DNA sequencing (mtDNA), single nucleotide polymorphisms (SNPs), and Y chromosome analysis. Autosomal ancestry informative markers (AIMs) are SNPs exhibiting highly contrasting allele frequencies between populations, thus enabling the inference of continental or regional origin. SNPs informative for ancestry also exist on mtDNA and the Y-chr and can place an individual into a haplogroup (Hg) which historically had restricted geographical distributions. Autosomal AIMs are inherited with recombination, representing input from both maternal and paternal lineages which can be useful in cases of ancestry admixture, while mtDNA and Y-chr SNPs infer the ancestry of either the maternal or paternal lineages alone. Biogeographical intelligence can provide valuable information for investigative leads when other avenues for biological profiling are exhausted (Fondevila *et al.* 2008; Phillips *et al.* 2009). Since determining country of origin is a critical part of the identification process of military remains, the recovery and analysis of a range of ancestry information, both genetic and physical is important for more reliable repatriation of unidentified servicemen.

The strength of any genetic profiling method for forensic testing relies on the comparison to population databases to estimate the frequency (or rarity) of seeing a particular DNA profile or haplogroup within a population, as a way to evaluate the ‘weight’ or confidence of results obtained (Steele & Balding 2014). Typically, a forensic database is created by generating DNA profiles from a random and representative sample of the population of interest. Australia has population databases for STR markers used in forensic investigation (Taylor *et al.* 2017), however no such databases relevant to an Australia population currently exist for either mtDNA or SNPs that are increasingly used in casework involving degraded DNA. Currently, mtDNA haplotype matches are evaluated against the EMPOP mtDNA population database (Parson & Dur 2007), which, at this time, does not include Australian samples, and

hence has limited use for Australian casework as it may not accurately capture the extent of genetic variation within the population. Similarly, SNPs that infer autosomal and Y-chr ancestry have also not been examined in a representative Australian population. The lack of such databases, both for the current and historical Australian population, limits the level of confidence that can be placed on assignment of country of origin and in an extreme case, could potentially risk erroneous repatriation.

To accurately determine using DNA technologies that remains recovered from historical battlefields belong to an Australian soldier, an understanding of the genetic ancestry components in the Australian population before the close of WWII in 1945 is needed. It is generally believed that all Australian servicemen serving across both world wars would carry exclusively European genetic ancestry because the Commonwealth Defence Act in 1909 ('Defence Act ' 1909) excluded any person not "substantially of European origin or descent" from enlisting with the Australian defence forces. However, the Aboriginal population of Australia prior to European arrival may have been as high as 750,000, and census records from 1911 and 1933 indicate a number of non-European ancestries present in Australia before 1945. In 1911 for example, 0.79% of males with non-European ancestry (full-blood) were recorded, with a further 0.17% recorded as 'half-caste' non-European ancestry in 1911 and 0.32% in 1933, indicating small amounts of genetic admixture (Australian Bureau of Statistics 1911, 1933). Notable migration events include the 1850's gold rush, where almost 12,000 Asian migrants arrived in Australia for mining (Gittins 1981) and rose up to approximately 37,000 by 1861 (Choi 1971), and the 'Afghan Cameleers' which saw approximately 2 - 3 thousand cameleers from India, Pakistan, Egypt, Persia, and Turkey arrive in Australia in the 1860's with working camels to develop and settle arid landscapes of Australia (Jones & Kenny 2007). During these times, interracial relationships occurred (Parkes 2009), also giving rise to descendants of admixed ancestry. Hence, the Australian population during the world wars could have been much more diverse in genetic ancestry than previously assumed, and those serving for Australia may have carried non-European ancestral DNA.

Considering the number of individuals of non-European origin residing in Australia before and during WWI and WWII, ancestry admixture could be a real possibility in those serving for Australia. Although the Defence Force did adopt a policy preventing those who were 'not substantially of European origin or descent' from enlisting (Defence Act 1909), there was no real way of knowing the extent of genetic admixture in individuals during this time,

especially since the race of enlistees was not recorded (Riseman 2013). Furthermore, a change of instructions in 1917 meant that an applicant with one parent of European origin was sufficient for enlistment (Military Order No. 200, 1917), with the aim to replace casualties lost during WWI. While by no means an exhaustive list, the WWII Nominal Roll (nominal-rolls.dva.gov.au/ww2), which records information regarding enlisted servicemen from WWII conflicts, including 'Country' and 'Place of Birth' outside Australia provides further possible evidence of ancestry admixture within the Australian wartime population (Table 1).

Country of Birth	Number of Servicemen Records
China	1044
Japan	55
Korea	8
Hong Kong	112
Macau	2
Thailand	1
Papua New Guinea	33
Africa	1398
India	1019

Table 1. Examples of servicemen records for enlisted individuals born outside of Australia and Europe from the WWII Nominal Roll.

Additionally, there are well documented instances where individuals with non-European heritage served for Australia, such as William 'Billy' Sing in WWI, a celebrated Sniper born to an English mother and Chinese father (Kennedy 2013). Captain Reginald Walter Saunders, commanding rifleman who served in the 2/7th battalion in New Guinea and his younger brother, also served in WWII along with the estimated 4000 other Aboriginal Australians across both world wars (Department of Veterans Affairs 2017; Australian War Memorial 2017b). Despite written evidence for non-European ancestry, the full extent of genetic admixture within those who served is largely unknown. As an extension, it is important to understand the ancestry components in the population to recognise the implications it may pose on biogeographic ancestry testing of recovered remains from historical conflicts.
The modern Australian population is likely to exhibit very different genetic composition than the early- to mid-1900s due to continued immigration since this time. Therefore, a population sample taken from the broader community at present would not accurately reflect frequency data in the population during WWI and WWII. There is a need for a representative population database to objectively evaluate the results obtained from forensic analysis of historical military remains.

In this study, the Australian Historical DNA Database (HADD), the first of its kind, was established as a reference database to allow for estimations of ancestry proportions in the Australian wartime population, using samples from members of the public who represent the Australian population before 1945. We investigated mtDNA ancestry by sequencing the mtDNA control region, and autosomal ancestry using a 31-plex SNP typing tool that distinguishes between five major continental ancestries. Data generated was collated into a frequency database to estimate the proportions of different ancestral groups in the Australian population during WWI and WWII. This database has the ability to complement and support current ancestry testing of degraded military remains recovered across historical battlefields, revealing which ancestries individuals could have carried in their DNA and ultimately improving the interpretations made from forensic testing of recovered remains.

Materials and Methods

Sample Collection

Over 800 samples were collected to represent the Australian population before the end of WWII in 1945, and the subsequent immigration from Eastern and Southern Europe (1945 onwards), South-East Asia (following the Vietnam War) and the Middle-East and Africa (1990's to present). Inclusion into the database required donors to be unrelated to one another (by checking genealogical information recorded by volunteers), born in Australia before 1945 or a direct descendant of people resident in Australia during this time (Supplementary File S1). DNA donors provided their year and place of birth, and the year and place of birth for their parents and grandparents (both maternal and paternal, if known).

Samples were self-collected via buccal swabs transferred onto FTA Indicating Micro cards. Ethical approval was granted by the University of Adelaide Human Research Ethics Committee (H-2015-120). All individuals gave written informed consent. Anonymity of the donors was preserved by use of unique laboratory codes.

DNA Extraction

DNA was extracted from a 5 mm^2 punch of FTA cards using the Chelex method (Walsh *et al.* 1991), QIAamp DNA Mini Kit (Qiagen, Hagen, Germany) or Charge Switch Forensic DNA Purification kit (Thermo Fisher Scientific Inc., Forster City, USA). Chelex extractions were

performed by washing the FTA card punch in 1 ml of DNA-free water, then transferring the punch to 200uL of 5% Chelex (Bio-Rad Laboratories, Hercules, USA), in 1X TE buffer in a 0.5 mL tube. The punch/Chelex solution was vortexed for 10 seconds, incubated at 100°C for 8 minutes, vortexed for 10 seconds and centrifuged at 20,000 rpm for 1 minute. A 150 uL aliquot of the supernatant was transferred to a 1.5 mL screw-cap tube. The QIAamp and Charge Switch extractions were performed according to manufacturer's instructions with the exception of a double elution of 75 uL of buffer for a total of 150 uL per sample.

mtDNA control region sequencing

Amplification of the mtDNA control region (16024-576) was performed in two separate PCR reactions targeting overlapping amplicons using primers L15996 (Vigilant *et al.* 1991) and H48 (unpublished) for amplicon one, and L16515 (unpublished) and H549 (Edson *et al.* 2004) for amplicon two (Table 2). The reverse primer for amplicon two was subsequently changed to H580 (Edson et al. 2004) to anneal outside the control region for full coverage. To facilitate direct sequencing of PCR products, forward primers were tagged with 16 nucleotides of the M13 forward sequence, and reverse primers were tagged with 17 nucleotides of the reverse M13 sequence.

Amplicon	Primer	Sequence (5' – 3')	Length
1	M13_L15996	**GTAAAACGACGGCCAG**CTCCACCATTAGCACCCAAAGC	
	M13_H48	**CAGGAAACAGCTATGAC**CCCCCAGACGAAAATACCAAATG	665
2	M13_L16515	**GTAAAACGACGGCCAG**ATCCGACATCTGGTTCCTACTTCA	
	M13_H549	**CAGGAAACAGCTATGAC**GGTGTATTTGGGGTTTGGTTG	646
	M13_H580	**CAGGAAACAGCTATGAC**TTGAGGAGGTAAGCTACATA	676

Table 2. PCR primer sequences used to amplify the mtDNA control region in two overlapping amplicons. Nucleotides in bold indicate M13 sequence. Amplicon length includes primer sequences.

PCR amplification was performed on 2 uL of DNA extract in a final reaction volume of 25 uL comprising 1x HiFi buffer, 2 mM $MgSO_4$, 250 nM of each dNTP, 200 nM of each primer, and 0.5 U of Platinum Taq DNA Polymerase High Fidelity (Thermo Fisher Scientific Inc., Forster City, USA). Thermocycling was done on a T1000 Thermal Cycler (Bio-Rad Laboratories, Hercules, USA) using the following conditions: 94°C for 2 min followed by 30 cycles of 94°C for 15 s, 55°C for 15 s, 68°C for 60 s and a final extension at 68°C for 10 min. Amplification success was assessed by gel electrophoresis on a 2 % agarose gel in 1x TBE buffer (100V for 45 min; Hyperladder IV DNA size ladder (Bioline Pty Ltd)). PCR product purification and double-strand sequencing reactions were performed at the Australian Genome Research Facility and electrophoresed on a 3730XL Genetic Analyser (Applied Biosystems, Forster City, USA), with 50 cm arrays and POP-7 polymer.

Sequence chromatograms were aligned to the revised Cambridge Reference Sequence (rCRS) (Andrews 1999) and evaluated using Geneious v 10.0.07 (Kearse *et al.* 2012) to obtain a consensus sequence. Sequence quality was manually evaluated and nomenclature adhered to the recommended guidelines (Parson & Bandelt 2007). Heteroplasmy was recorded if a second peak was detected at a position with at least 25% fluorescence of the major peak. All subsequent analyses were performed using the entire mtDNA control region (16024-576), excluding length variation at positions 309, 515-522, 573 and 16193 for haplotype calling, and mutational hotspot variants 16182, 16183, or 16519 for haplogroup assignment.

Haplogroup affiliation was performed by entering haplotypes into HaploGrep2 (Weissensteiner *et al.* 2016) based on PhyloTree Build 17 (van Oven 2015), in parallel with EMPOP (Parson & Dur 2007) and mtDNAmanager (Lee *et al.* 2008), with some being checked manually using PhyloTree Build 17. Samples were assigned a specific sub-haplogroup (e.g. H6a1a), haplogroup (e.g. H6a) and one of 28 broad haplogroups (e.g. H) as described in Figure 1, according to the PhyloTree phylogeny (refer to Supplementary File S2). Haplogroup K was treated as a separate broad haplogroup, despite branching off from haplogroup U8, likewise for haplogroup V, which resides in broad haplogroup HV0, due to the representation of these subgroups in this dataset. Assigning the most likely maternal ancestry and continental affiliation for each sample was performed through an extensive survey of the literature for the origins and geographical dispersion of mtDNA haplogroups (van Oven *et al.* 2011a), including EMPOP (Parson & Dur 2007).

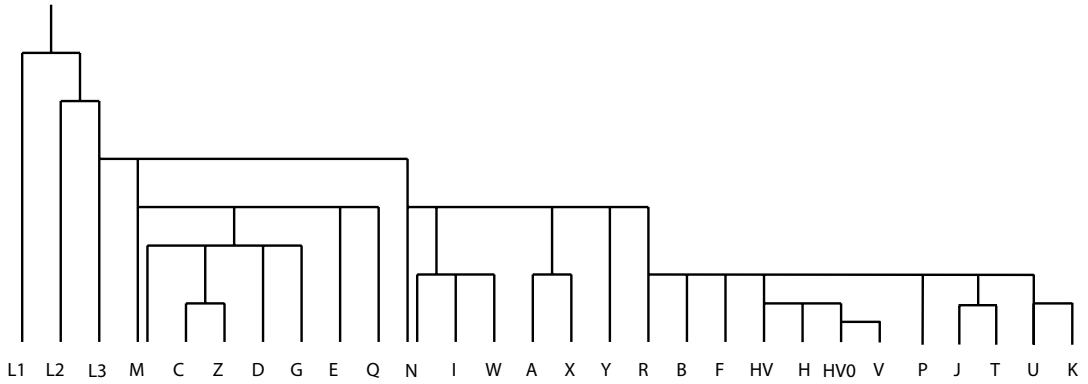


Figure 1. Simplified mtDNA haplogroup tree. Haplogroups listed above were considered as 'broad' when assigning samples.

Autosomal SNP analysis

Thirty-one autosomal ancestry-informative SNPs covering five global populations (Africa = AFR, Europe = EUR, East Asia = EAS, Native America = AMR and Oceania = OCE) were genotyped using the Global AIMs Nano set (de la Puente et al. 2016), and SNaPshot multiplex chemistry as described in de la Puente et al. 2016. The Global AIMs Nano set was selected for the ability to differentiate between five major geographical populations, including Oceania which is applicable for an Australian population. Capillary electrophoresis was performed on a 3500 Genetic Analyzer (Applied Biosystems, Forster City, USA) with 36 cm arrays and POP-4 polymer. Electropherograms were analysed for genotype calling in GeneMapper ID version 5.0 (Applied Biosystems, Forster City, USA) using a custom panel and bin settings.

For assignment of biogeographic ancestry, each sample genotype was compared to a reference population set comprising genotypes from 402 individuals from African (AFR, n = 108), East Asian (EAS, n = 103), European (EUR, n = 99), Native American (AMR, n = 64), and Oceanian (OCE, n = 28) populations. Reference population genotypes were obtained from the 1000 Genomes Phase II (The 1000 Genomes Project Consortium 2015) and Stanford University HGDP-CEPH (Cann *et al.* 2002) datasets, and were carefully selected from populations that show minimal admixture (Supplementary File S3). Ancestry assignment was performed using the Bayesian classifier, Snipper v2.5 (mathgene.usc.es/snipper/), with Hardy-Weinberg principle applied. Likelihood ratios (LR) for ancestry classifications were used for direct ancestry estimation, and principle component analysis (PCA) was performed in RStudio (v1.1.442) using the *SNPassoc* package (Gonzalez *et al.* 2007) to visually summarise the genetic differences and similarities of the sample genotypes to the reference populations.

The level of admixture for each sample was assessed further by applying the admixture model to the autosomal SNP data in STRUCTURE v.2.3.4 (Porras-Hurtado *et al.* 2013). The reference population set described above was used for population membership analysis of the database samples in STRUCTURE. K:5 has been previously identified as the optimum number of clusters for the Global AIMs Nano set reference populations (de la Puente et al. 2016) and was used for this analysis. Analyses with STRUCTURE were performed using the following parameters: five iterations (for K=5) of 100,000 burnin steps and 100,000 MCMC steps, correlated allele frequencies under the Admixture model (including POPFLAG).

Estimated membership coefficients from STRUCTURE analysis were used to construct population membership bar plots with CLUMPAK v.1.1 (Kopelman *et al.* 2015).

Through this analysis, it was found that there were a number of underperforming SNPs, namely rs12498138, rs9908046 and rs2069945. To test if this affected population assignment, the data was analysed with and without these markers (including samples for which these were successfully genotyped) as discussed above.

Statistical Analyses

As the database established in this study represents only a subset of the target population, confidence intervals for the frequency of non-European mtDNA and autosomal ancestry classifications were conducted as an estimate for the true population parameter (and to account for sampling variation). Calculation of confidence intervals were performed based on the frequency point estimate in the database using the Wilson interval (Wilson 1927) to provide more conservative upper limits an avoid underestimating non-European ancestries in the population.

Quality Control

A positive control of known genotype was included in all PCRs, sequencing reactions, and SNP typing to ensure reproducibility and accuracy between batches. Extraction blank controls and PCR negative controls were also included to monitor for contamination, allele and locus drop-in and other artefacts. To ensure high quality mtDNA control region data, each site was required to be covered by two sequence reads, and to have two independent reviewers in agreement of consensus calling.

Results

Sample Database

Over 800 samples were collected between May 2016 and September 2017. A subset of two hundred and fifty-nine randomly selected samples were analysed as part of this study.

mtDNA

All samples in the database obtained full sequences from 16024-548, 189 of which had further coverage to position 576 following a change to the H580 primer. A total of 222 haplotypes were detected among the 259 samples, 197 (76.1%) of which were unique sample haplotypes. Of the remaining 23.9% of haplotypes, 25 were shared by more than one sample. The two most common haplotypes (16519C, 263G and 16519C, 152C, 263G) both appeared in five individuals each (1.9%). The second most common haplotype (16298C, 72C, 263G) appeared in four samples (1.5%).

Eleven instances of point heteroplasmy were observed in 10 samples (3.8% of samples), at nine different sites. Six were Y transitions, and three were R transitions (189R, 195Y, 215R, 16092Y, 16093Y, 16111Y, 16188Y, 16192Y, 16227R). The most frequent, 16093Y and 195Y, appeared in two samples each. All were consistent with previous studies and have been seen in the EMPOP database with the exception of 16227R. Heteroplasmic positions were not considered for haplogroup assignment.

The 197 mtDNA haplotypes were assigned to 72 haplogroups (Table 3, Column 2) within 15 broad haplogroup classifications (Table 3, Column 1). Details of the haplotypes, specific sub-haplogroups, haplogroups and broad haplogroup assignment can be found in (Supplementary File S2). The most common haplogroup (H*) was found in 30 individuals, followed by J1c (17 individuals) and U5a (15 individuals).

Broad Hg	Hg	No. of Samples	%
N		**1**	**0.4**
	N1a	1	0.4
I		**3**	**1.2**
	I1a	2	0.8
	I2	1	0.4
W		**6**	**2.3**
	W*	3	1.2
	W1	1	0.4
	W5	1	0.4
	W6	1	0.4
X		**2**	**0.8**
	X2	2	0.8
S		**1**	**0.4**
	S1	1	0.4
R		**4**	**1.5**
	R*	3	1.2
	R1a	1	0.4
B		**1**	**0.4**
	B2d	1	0.4
U		**35**	**13.5**
	U2	2	0.8
	U3	1	0.4
	U4	6	2.3
	U5a	15	5.8
	U5b	8	3.1
	U6	2	0.8
	U8	1	0.4

Broad Hg	Hg	No. of Samples	%
K		**21**	**8.1**
	K1	18	6.9
	K2	3	1.2
J		**33**	**12.7**
	J1	3	1.2
	J1b	11	4.2
	J1c	17	6.6
	J2a	2	0.8
T		**28**	**10.8**
	T*	1	0.4
	T1a	7	2.7
	T2	2	0.8
	T2b	10	3.9
	T2c	2	0.8
	T2e	1	0.4
	T2f	4	1.5
	T2g	1	0.4
HV		**5**	**1.9**
	HV*	1	0.4
	HV1	1	0.4
	HV8	1	0.4
	HV9	1	0.4
	HV15	1	0.4
HV0		**10**	**3.9**
	HV0*	10	3.9
V		**3**	**1.2**
	V*	1	0.4
	V6	1	0.4
	V16	1	0.4

Broad Hg	Hg	No. of Samples	%
H		**106**	**40.9**
	H*	30	11.6
	H1	3	1.2
	H1a	9	3.5
	H1b	7	2.7
	H1c	8	3.1
	H1e	3	1.2
	H1i	1	0.4
	H1o	2	0.8
	H1q	2	0.8
	H2a	11	4.2
	H3b	1	0.4
	H3h	1	0.4
	H3v	2	0.8
	H3z	1	0.4
	H4	1	0.4
	H5	5	1.9
	H5a	1	0.4
	H5b	2	0.8
	H6a	3	1.2
	H7a	1	0.4
	H7d	1	0.4
	H7h	2	0.8
	H10e	1	0.4
	H11	2	0.8
	H11a	1	0.4
	H16b	1	0.4
	H20	1	0.4
	H24	1	0.4
	H39	1	0.4
	H85	1	0.4
Total		**259**	**100**

Table 3. Frequencies of the mtDNA haplogroups and broad haplogroups in 259 HADD samples. * Denotes samples which could not be resolved into further haplogroups. Sub-haplogroup details can be found in Supplementary File S2. Hg = haplogroup.

Forty-nine samples could only be assigned to broad haplogroups, due to insufficient variation in the control region to allow deeper resolution in these cases (30 samples assigned to H, ten to HV0, three to R, three to W, one to HV, one to V, and one to T).

The most frequent broad haplogroup in the dataset was H (40.9%). The most frequent haplogroups within H were H1 (13.5%) and H2 (4.2%). Broad haplogroup HV0 accounts for 3.9% of the samples, with less than half represented by haplogroup V (1.2%). Ancestral to H, haplogroup HV accounted for 1.9% of the dataset. Following haplogroup H, haplogroup U (encompassing U2, U3, U4, U5, U6 and U8 haplotypes) accounts for 13.5% of the database and is the second highest contributor to the dataset. Haplogroup K was treated as a separate group and was present in 8.1% of the samples. Haplogroup J and T defines 12.7% and 10.8% of the dataset respectively and are divided into J1 (11.9%), J2 (0.8%), T* (0.4%), T1 (2.7%) and T2 (7.8%) subclades. One sample lacked further variation and could only be assigned to the basal branch of haplogroup T. Clade N is present in 5.1% of the samples, represented by haplogroups W (2.3%), I (1.2%), X (0.8%), N1 (0.4%), and S (0.4%).

Autosomal SNPs

Thirty-five samples produced a full 31-locus SNP genotype, with the remaining 224 samples producing partial profiles (average 28 out of 31 loci). SNPs rs12498138 (AMR informative), rs9908046 (OCE informative), and rs2069945 (tri-allelic) were the three poorest performing loci and were unable to be genotyped in 163, 85, and 82 samples respectively. Despite allelic dropout, final population assignment was not affected, and reliable ancestry estimation was able to be achieved for all samples.

All samples in the dataset were predicted as European ancestry using Snipper, with probabilities at least 1 billion times more likely European than any of the other four continental populations. No ancestry admixture was detected in any of the samples using Snipper, with all profiles returning a ‘100% EUR’ classification. The PCA plot for all samples compared to 402 reference population genotypes detailed in de la Puente et al. 2016 show a clustering of the database samples around the European reference population. The effect of underperforming SNPs was tested by removal of rs12498138, rs9908046, and rs2069945from the dataset. All probabilities and admixture prediction values in Snipper were maintained (1 billion times more likely EUR than any of the other populations, and 100% EUR), with the resulting PCA showing minimal difference to the original plot including the problematic loci (Figure 2).

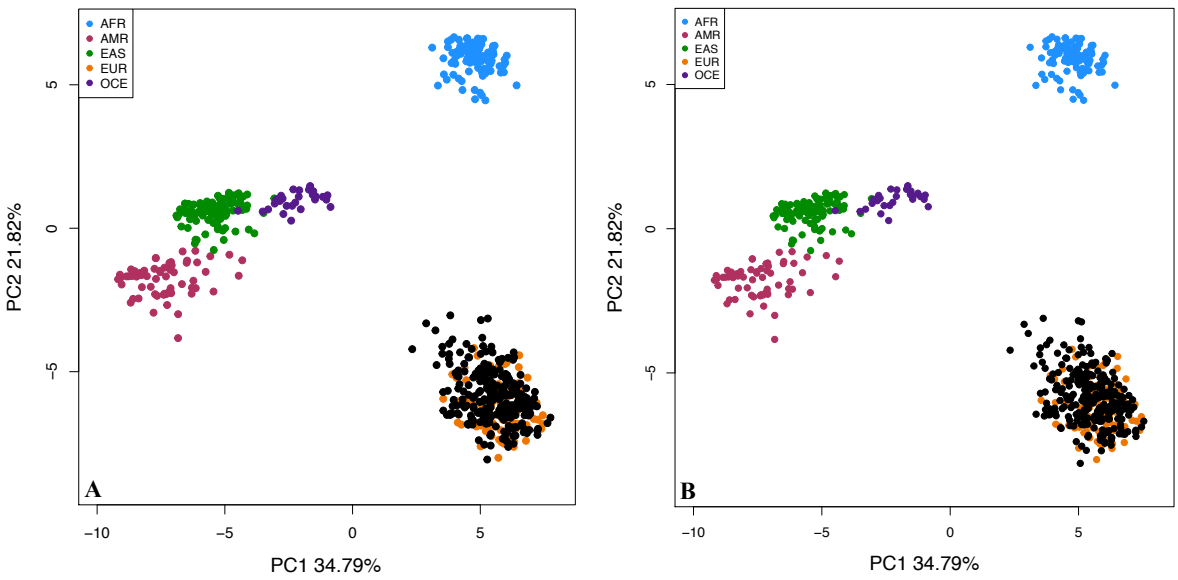


Figure 2. PC1 vs PC2 from PCA of database samples (black) against 402 genotypes across five reference population groups (A), and when removing underperforming SNPs (B) (rs12498138, rs9908046, and rs2069945). AFR = blue, AMR = red, EAS = green, EUR = orange, OCE = purple. Remaining components are plotted in Supplementary File S4.

Further analysis in STRUCTURE (K = 5) showed all samples had the highest population membership to the EUR reference population (Figure 3). Overall, the database samples showed a 97.2% cluster membership to the EUR reference population, with less than 1% membership to each of other four populations. The bar plot visualised in CLUMPAK shows similar patterns of population assignment of the database samples, with all samples having predominantly European ancestry, with minor to no ancestry components from other population groups. The membership proportion for each sample can be found in Supplementary File S5.

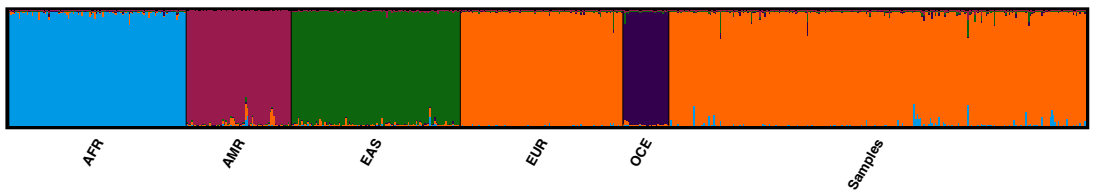


Figure 3. CLUMPAK bar plot of STRUCTURE analysis (K=5), showing membership proportions of each sample to the reference population groups. Each bar represents one individual sample.

Non-European ancestry components in the database were further evaluated using the ancestry membership coefficients obtained from the STRUCTURE analysis and were plotted against the European membership coefficients for each sample (Figure 4).

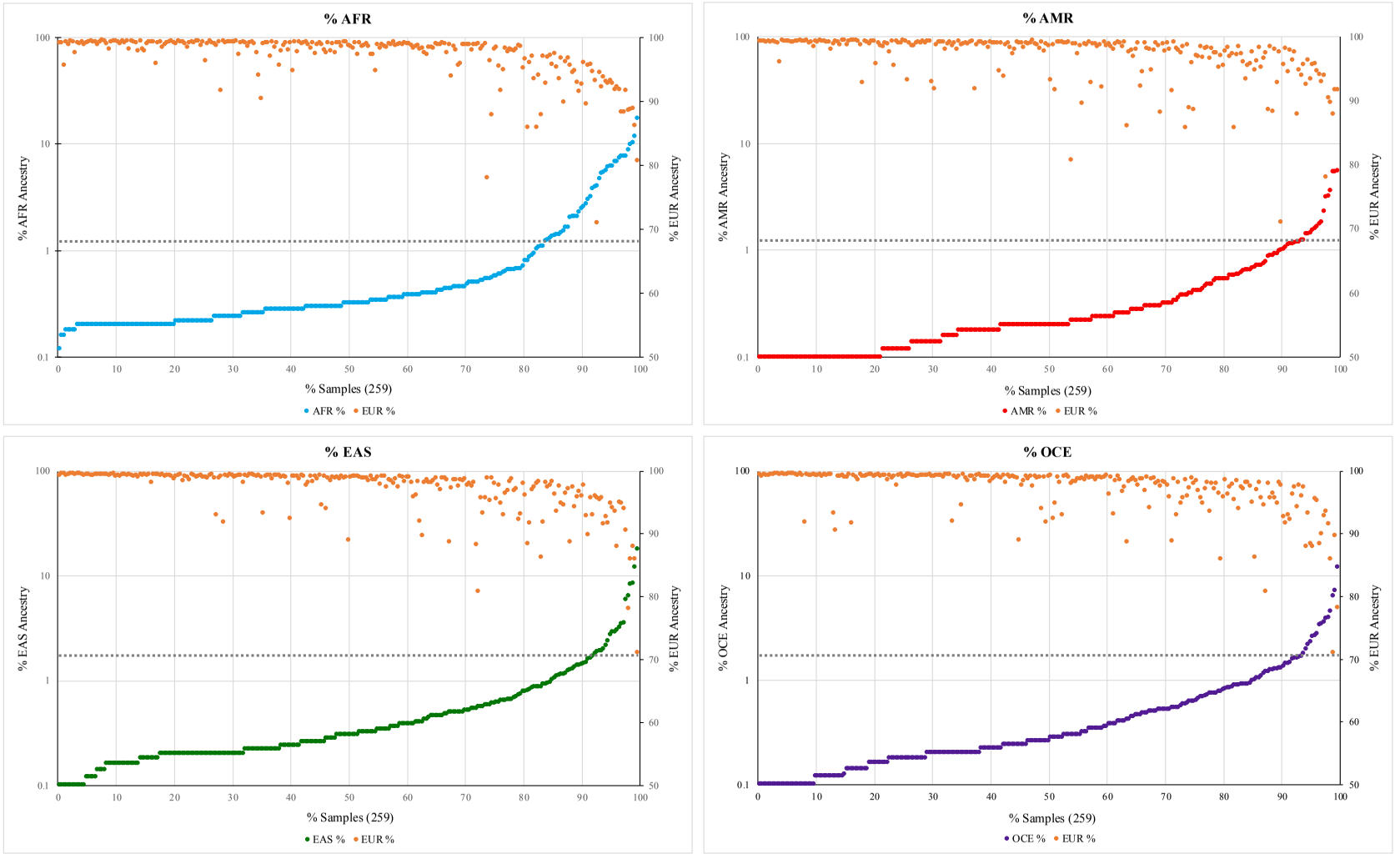


Figure 4. Non-European ancestry proportions (left hand Y-axis, log scale) compared with European ancestry proportion (right hand Y-axis, orange) for each sample in the HADD, ordered by % non-European ancestry. AFR = African (blue); AMR = Native American (red); EAS = East Asian (green); OCE = Oceanian (purple); EUR = European (orange). Horizontal dotted line indicates 1% non-European ancestry threshold.

Across all samples, 82.2%, 89.9%, 84.9% and 85.7% had an AFR, AMR, EAS and OCE ancestry proportion below 1%, respectively. The highest proportion of AFR ancestry (17.2%) was seen in a sample with 80.8% EUR ancestry. The highest proportion of AMR ancestry (5.5%) was seen in a sample with 91.7% EUR ancestry. The highest proportion of EAS ancestry (17.6%) was seen in a sample with 71.2% EUR ancestry. The highest proportion of OCE ancestry (11.8%) was seen in a sample having 78.3% EUR ancestry. The two samples which indicated Australian Aboriginal and Native American maternal ancestry (mtDNA) showed predominantly European autosomal ancestry membership proportions (Table 4).

Sample	mtDNA Hg	% AFR	% AMR	% EAS	% EUR	% OCE
157	S1	1.54	3.38	1.6	89.84	3.64
238	B2d	0.24	0.72	0.38	98.12	0.54

Table 4. Population membership proportions of two samples that revealed non-European mtDNA haplogroups

Statistical Analysis

For mtDNA, two samples carried non-European haplogroups, namely Australian Aboriginal S1and Native American founder lineage B2d. Wilson confidence intervals were calculated for each using a point estimate of 0.4%. The 95% confidence interval for the proportion of Australian Aboriginal mtDNA ancestry, and Native American mtDNA ancestry in our dataset is 0.07% - 2.15% each. Therefore, it can be said with 95% confidence that the true proportion of individuals living in Australia before 1945 carrying non-European mtDNA haplogroups (based on Australian Aboriginal and Native American combined) is between 0.2% - 2.8%.

For autosomal ancestry, no sample with predominantly non-European ancestry was predicted by Snipper, PCA and STRUCTURE. However, confidence intervals were still calculated based on this point estimate due to database sample size. The 95% confidence interval for the proportion of non-European autosomal ancestry in the Australian population before 1945 is 0% - 1.5%. Therefore, it can be said with 95% confidence that the true proportion of individuals living in Australia before 1945 carrying predominantly non-European autosomal ancestry is between 0% - 1.5%.

Discussion

To facilitate forensic ancestry determination for repatriation and identification of Australian historical remains we have created a population database, the Historical Australian DNA

Database (HADD). This database records mtDNA control region haplotypes (and haplogroup), and 31-locus autosomal ancestry SNP genotypes, allowing estimation of the genetic ancestry proportions in an early-to-mid 1900's Australian population. Currently, 259 individuals have been analysed but to date over 800 samples have been collected. The maternal inheritance of mtDNA presents concerns over estimating biogeographic ancestry when considered alone (Phillips *et al.* 2009; Phillips 2015). For this reason, we chose to combine mtDNA control region data with autosomal ancestry informative SNP data to allow improved resolution of ancestry and to detect possible admixture. No suitable forensic population databases currently exist for differentiation of human remains such as those recovered from historical battlefields that may be Australian. The HADD presents the first Australian ancestry informative dataset that can be utilised for forensic purposes.

Analysis of mtDNA revealed that haplogroups in the database are largely typical of West Eurasia, as reflected in mtDNA data previously published (Lutz *et al.* 1998; Richards *et al.* 2000; Achilli *et al.* 2007; Brandstatter *et al.* 2007; Hedman *et al.* 2007; Richard *et al.* 2007; Irwin *et al.* 2008; Turchi *et al.* 2008; Turchi *et al.* 2016), with the exception of one sample belonging to the Native American-specific haplogroup B2d (Bandelt *et al.* 2003; van Oven *et al.* 2011b), and one sample belonging to the Australian Aboriginal-specific haplogroup S1 (Nagle *et al.* 2017; Tobler *et al.* 2017). Similar to European and Near Eastern populations (Richards *et al.* 2000; Pereira *et al.* 2005; Roostalu *et al.* 2007; Chaitanya *et al.* 2016), the most frequent haplogroup in the HADD is haplogroup H. The highest contributors to this group were subgroups H1 and H2. H1 is one of the most common subclades of H seen in Western Europe and Morocco, peaking in Iberia (Hernández *et al.* 2017). H2, more specifically H2a, occurs more frequently in Eastern Europe, extending to Central Asia in low frequencies (Loogvali *et al.* 2004).

Haplogroup U, encompassing subgroups U2, U3, U3, U4, U5, U6 and U8 in this dataset, makes up the second largest proportion of samples. Subclade U5, the most ancient European mtDNA haplogroup (Malyarchuk *et al.* 2010), represents more than half of samples belonging to haplogroup U, and has a wide distribution over Western Eurasia and South Asia (van Oven *et al.* 2011b). U6, represented by two samples is of North African origin (Rando *et al.* 1998) and appears in West Mediterranean populations <7% (Plaza *et al.* 2003) representing recent gene-flow from Northern Africa (Hervella *et al.* 2016). Haplogroup K (within U8) represents a small portion of the dataset. It peaks in frequency in France,

Hungary and the British Isles, however still exists in relatively low frequencies globally (García *et al.* 2011; Turchi *et al.* 2016).

The majority of samples in haplogroup J fell within J1c, representing half of the total J1 lineages in this dataset, the rest assigned to J1b. J1 has a predominantly central European distribution (Turchi et al. 2016). J2 is present at a much lower frequency, represented by J2a1 subgroups. Haplogroup T is present in comparable frequencies to populations over Europe (Chaitanya et al. 2016; Turchi et al. 2016), and falls into two distinct subclades in the dataset, T1and T2. T2b subgroups represent over half of the T2 individuals and is also predominantly European in distribution (Pala *et al.* 2012).

Haplogroup W was detected in a small subset of samples and exists in relatively low frequencies across the European continent, the Near East and West Asia. It peaks across Eastern Europe (~6.5% of population), India (~6% of population), and Northern Europe (~4% of population) (Hedman *et al.* 2007; Olivieri *et al.* 2013). The remaining N groups (I, X, N1) occur at relatively low frequency in the dataset.

Two samples revealed non-European mtDNA lineages, one assigned to B2d (Native American specific), and one assigned to S1 (Australian Aboriginal specific). The autosomal data for these samples indicated clustering with the European reference population with a strong likelihood (> 1 billion times more likely) in Snipper and the PCA plot, whereas STRUCTURE detected very minimal amounts of membership to other populations (0.68% AFR, 0.5% AMR, 0.38% EAS, 0.24% OCE versus 98.1% EUR ancestry for the B2d sample, and 1.54% AFR, 3.4% AMR, 1.6% EAS, 3.64% OCE ancestry versus 89.9% EUR for the S1 sample). Haplogroup S1 is not entirely unexpected considering the presence of Aboriginal Australians in Australia for approximately 50,000 years (Tobler *et al.* 2017), and the lack of Oceanian autosomal SNPs indicates an admixed individual with Aboriginal Australian maternal ancestry and European autosomal ancestry. The occurrence of European autosomal ancestry has been observed in Aboriginal Australian samples in a previous study (Santos *et al.* 2016), and represents genetic admixture in the Australian population post-European settlement. The Native American maternal lineage with a European autosomal ancestry may reflect admixture before or after migration to Australia.

Forty-nine samples could not be assigned into specific haplogroups due to a lack of haplogroup defining polymorphisms in the mtDNA control region. Considering the wealth of

variation (~70%) present in the coding region (Brotherton et al. 2013; Ma et al. 2018), interrogation of diagnostic SNPs from outside of the control region may improve resolution into further sub-haplogroups for more specific geographical assignment. This may be performed either through SNaPshot-based SNP multiplexes designed to target a small number of coding region SNPs as an independent marker set to resolve haplogroups (as demonstrated previously in (Corach *et al.* 2010), or by the use of whole mitochondrial genome sequencing to capture a greater level of genetic variation for finer resolution into sub-haplogroups (Chaitanya *et al.* 2015; Morales-Arce *et al.* 2017). The typing of coding region SNPs may be a beneficial and economic addition to the database in the future for improved resolution and maternal biogeographic ancestry assignment.

For the autosomal ancestry SNPs, three underperforming loci resulted in a high proportion of partial profiles, however the biogeographic ancestries of all samples were still able to be predicted with confidence and high statistical power. Differences in the number of alleles successfully typed has previously been observed in inter-laboratory studies using SNaPshot based SNP panels (Musgrave-Brown *et al.* 2007). One of the underperforming SNPs (rs2069945), is a tri-allelic marker included for contamination detection and is not ancestry informative for a specific population (de la Puente et al. 2016). To determine if the dropout of these SNPs had an influence on ancestry prediction, the data was re-analysed following the removal of the three underperforming markers. Results showed that locus dropout had minimal effect on ancestry classification. Furthermore, the Global AIMs Nano has been demonstrated to achieve 100% ancestry assignment success even when excluding the 14 most informative SNPs (i.e. the SNPs with the highest divergence) (de la Puente et al. 2016), demonstrating the panel maintains a high level of informativeness even when analysing partial profiles.

While mtDNA haplogroup frequencies revealed that there may have been a small contribution of non-European ancestry in the maternal lineages, autosomal SNP analysis predicted samples as predominantly European ancestry. Confidence intervals estimate that the true proportion of non-European ancestry could range from 0 - 2.8% of the Australian population before 1945 (including both mtDNA ancestry and autosomal ancestry frequencies). These confidence intervals allow investigators to objectively evaluate the evidential weight that can be placed on ancestry results and demonstrates the possibility that Australian WWI and WWII servicemen may not have exclusively carried European ancestry. However, it is important to stress the limitations of the database at this time with only 259

samples typed out of at least 6.6 million individuals living in Australia in the time period (from census data for 1933). Confidence intervals are strongly dependant on sample size but were calculated for this study regardless of size limitations and should be interpreted with caution. A larger database is needed to properly evaluate frequency estimates of non-European genetic ancestry, and thus the potential implications on ancestry testing. Nonetheless, this study presents the first step towards achieving this outcome.

The different ancestry signals observed in two samples between the mtDNA ancestry and autosomal SNPs further demonstrates and supports the importance of including autosomal and uniparental (mtDNA and Y chromosome) makers to avoid misreporting an individual's overall biogeographic ancestry when one genetic target is considered in isolation (Phillips *et al.* 2009; Corach *et al.* 2010; Lao *et al.* 2010; Prestes *et al.* 2016). This also becomes important when considering that the wide distribution of some basal mtDNA haplogroups (which may be assigned to samples which do not display sufficient diagnostic SNPs in the control region) may not allow for confident classification to one geographical region, and other targets can be analysed to improve assessments of ancestry. While not analysed in this study, it is recommended that analysis of Y chromosome markers be performed for male samples in the database for a more comprehensive survey of the maternal, paternal and autosomal biogeographic ancestry proportions in the Australian population. Based on the recommendation of combining different marker types for ancestry inference, another challenge may arise that requires the conception of a suitable statistical process to provide a single probability which encompasses results from all marker types tested, demanding unanimity among a board of both practitioners and forensic statisticians.

This study presents the first step towards creating a suitable population database to assist in understanding the probability of different ancestry groups in the Australian population before 1945. A larger sample size with the addition of Y chromosome analysis, will increase the knowledge of genetic ancestry in the Australian population during this time. This will improve genetic ancestry testing and assignment of country-of-origin to unknown remains which could belong to an Australian servicemen or woman. In addition to informing DNA analysis methods for ancestry determination of historical remains, the Historical Australian DNA Database may also provide a foundation for future studies into the ancestry group distributions within the modern Australian population. The HADD also has a potential application to broader forensic ancestry examination of multiple individuals such as in open international mass disasters and missing persons cases where the sorting of people into

ancestry groups is analogous to grouping people based on biological sex to help guide investigations.

Conclusion

High quality control-region mtDNA sequences and 31-plex autosomal ancestry SNP profiles were generated from 259 individuals to form the core of the Historical Australian DNA Database (HADD), with the aim of identifying genetic ancestry variation in the Australian population before the end of WWII. The pattern of mtDNA variation and autosomal SNP profiles is characterised by an overall high frequency of predominantly European genetic ancestry. However, mtDNA haplogroup composition reveals a low frequency of genetic ancestry from non-European origins, reflecting Australia's known demographic history, and further advocates for the use of a multi-gene approach for the estimation of biogeographic ancestry. Future efforts will be placed on increasing the database size and generating Y-chromosome data for a more robust representation of the genetic composition of the Australian population before 1945. Nonetheless, this approach has generated the foundations of the first historical DNA database for Australia using multiple genetic targets and serves as a growing population database for which to evaluate ancestry determination results from Australian historical remains.

Competing Interests

The authors declare no competing interests

Acknowledgements

We are grateful for members of the Australian public for generously donating a DNA sample to assist in the project. We thank Leanne van Weert, Jennifer Young and Kelly Hill for assistance in sample processing; Adrian Linacre (Flinders University) and Forensic Science South Australia for providing access to equipment.

References

Achilli, A., Olivieri, A., Pala, M., Metspalu, E., Fornarino, S., Battaglia, V., Accetturo, M., Kutuev, I., Khusnutdinova, E., Pennarun, E., et al. 2007. Mitochondrial DNA Variation of Modern Tuscans Supports the Near Eastern Origin of Etruscans, *Am J Hum Genet*, 80, 759-68.

Andrews, R.M., Kubacka, I., Chinnery, P.F., Lightowlers, R.N., Turnbull, D.M., Howell, N. 1999. Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA, *Nature Genet*, 23, 147.

Australian Bureau of Statistics 1911, *Census of the Commonwealth of Australia, Vol II- Part II Birthplaces*, viewed November 17 2017, <http://www.abs.gov.au/AUSSTATS/abs@.nsf/DetailsPage/2112.01911?OpenDocument >.

Australian Bureau of Statistics 1933, *Census of the Commonwealth of Australia, Vol II- Part X Birthplace*, viewed November 17 2017, <http://www.abs.gov.au/AUSSTATS/abs@.nsf/DetailsPage/2110.01933?OpenDocument >.

Australian War Memorial 2017a, *Deaths as a result of service with Australian units*, viewed November 13 2017, <https://www.awm.gov.au/articles/encyclopedia/war_casualties>.

Australian War Memorial 2017b, *Indigenous defence service*, viewed December 9 2017, <www.awm.gov.au/articles/encyclopedia/indigenous>.

Bandelt, H.J., Herrnstadt, C., Yao, Y.G., Kong, Q.P., Kivisild, T., Rengo, C., Scozzari, R., Richards, M., Villems, R., Macaulay, V., et al. 2003. Identification of Native American founder mtDNAs through the analysis of complete mtDNA sequences: some caveats, *Ann Hum Genet*, 67, 512-24.

Brandstatter, A., Niederstatter, H., Pavlic, M., Grubwieser, P. & Parson, W. 2007. Generating population data for the EMPOP database - an overview of the mtDNA sequencing and data evaluation processes considering 273 Austrian control region sequences as example, *Forensic Sci Int*, 166, 164-75.

Cann, H.M., de Toma, C., Cazes, L., Legrand, M.F., Morel, V., Piouffre, L., Bodmer, J., Bodmer, W.F., Bonne-Tamir, B., Cambon-Thomsen, A., et al. 2002. A human genome diversity cell line panel, *Science*, 296, 261-2.

Chaitanya, L., Ralf, A., van Oven, M., Kupiec, T., Chang, J., Lagacé, R. & Kayser, M. 2015. Simultaneous Whole Mitochondrial Genome Sequencing with Short Overlapping Amplicons Suitable for Degraded DNA Using the Ion Torrent Personal Genome Machine, *Human Mutat*, 36, 1236-47.

Chaitanya, L., van Oven, M., Brauer, S., Zimmermann, B., Huber, G., Xavier, C., Parson, W., de Knijff, P. & Kayser, M. 2016. High-quality mtDNA control region sequences from 680 individuals sampled across the Netherlands to establish a national forensic mtDNA reference database, *Forensic Sci Int Genet*, 21, 158-67.

Choi, C.Y. 1971, 'Chinese immigration and settlement in Australia, with special reference to the Chinese in Melbourne', School of Social Science, Doctor of Philosophy thesis, Australian National University.

Church, M. 1995. Determination of race from the skeleton through forensic anthropological methods, *Forensic Sci Rev*, 7, 1-39.

Corach, D., Lao, O., Bobillo, C., Van Der Gaag, K., Zuniga, S., Vermeulen, M., Van Duijn, K., Goedbloed, M., Vallone, P.M., Parson, W., et al. 2010. Inferring Continental Ancestry of Argentineans from Autosomal, Y-Chromosomal and Mitochondrial DNA, *Ann Hum Genet*, 74, 65-76.

de la Puente, M., Santos, C., Fondevila, M., Manzo, L., Carracedo, A., Lareu, M.V. & Phillips, C. 2016. The Global AIMs Nano set: A 31-plex SNaPshot assay of ancestry-informative SNPs, *Forensic Sci Int Genet*, 22, 81-8.

'Defence Act ' 1909, Australia.

Department of Veterans Affairs 2016, *Office of Australian War Graves*, viewed November 10 2017, <https://www.dva.gov.au/commemorations-memorials-and-war-graves/office-australian-war-graves/about-office-australian-war. >.

Department of Veterans Affairs 2017, *Indigenous Australians at War*, viewed December 9 2017, <www.dva.gov.au/i-am/aboriginal-andor-torres-strait-islander/indigenous-australians-war>.

Edgar, H.J. 2013. Estimation of ancestry using dental morphological characteristics, *J Forensic Sci*, 58 Suppl 1, S3-8.

Edson, S.M., Ross, J.P., Coble, M.D., Parsons, T.J. & Barritt, S.M. 2004. Naming the Dead — Confronting the Realities of Rapid Identification of Degraded Skeletal Remains, *Forensic Sci Rev*, 16, 63.

Fondevila, M., Phillips, C., Naveran, N., Fernandez, L., Cerezo, M., Salas, A., Carracedo, A. & Lareu, M.V. 2008. Case report: identification of skeletal remains using short-amplicon marker analysis of severely degraded DNA extracted from a decomposed and charred femur, *Forensic Sci Int Genet*, 2, 212-8.

García, O., Fregel, R., Larruga, J.M., Álvarez, V., Yurrebaso, I., Cabrera, V.M. & González, A.M. 2011. Using mitochondrial DNA to test the hypothesis of a European post-glacial human recolonization from the Franco-Cantabrian refuge, *Heredity*, 106, 37-45.

Gittins, J. 1981. *The diggers from China: The story of Chinese on the goldfields*, Quartet Books Australia, Melbourne, Australia.

Gonzalez, J.R., Armengol, L., Sole, X., Guino, E., Mercader, J.M., Estivill, X. & Moreno, V. 2007. SNPassoc: an R package to perform whole genome association studies, *Bioinformatics*, 23, 644-5.

Hedman, M., Brandstatter, A., Pimenoff, V., Sistonen, P., Palo, J.U., Parson, W. & Sajantila, A. 2007. Finnish mitochondrial DNA HVS-I and HVS-II population data, *Forensic Sci Int*, 172, 171-8.

Hernández, C.L., Dugoujon, J.M., Novelletto, A., Rodríguez, J.N., Cuesta, P. & Calderón, R. 2017. The distribution of mitochondrial DNA haplogroup H in southern Iberia indicates ancient human genetic exchanges along the western edge of the Mediterranean, *BMC Genet*, 18, 46.

Hervella, M., Svensson, E.M., Alberdi, A., Günther, T., Izagirre, N., Munters, A.R., Alonso, S., Ioana, M., Ridiche, F., Soficaru, A., et al. 2016. The mitogenome of a 35,000-year-old Homo sapiens from Europe supports a Palaeolithic back-migration to Africa, *Sci Rep*, 6, 25501.

Irwin, J., Saunier, J., Strouss, K., Paintner, C., Diegoli, T., Sturk, K., Kovatsi, L., Brandstätter, A., Cariolou, M.A., Parson, W., et al. 2008. Mitochondrial control region sequences from northern Greece and Greek Cypriots, *Int J Legal Med*, 122, 87-9.

Jones, P. & Kenny, A. 2007. *Australia's Muslim cameleers : pioneers of the inland, 1860s-1930s / Philip Jones and Anna Kenny*, ed. A Kenny Wakefield Press, Kent Town, S. Aust.

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper, A., Markowitz, S., Duran, C., et al. 2012. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data, *Bioinformatics*, 28, 1647-9.

Kennedy, A. 2013. *Chinese Anzacs : Australians of Chinese descent in the defence forces 1885-1919 : revised to include New Zealand-born Chinese of the New Zealand Expeditionary Force 1914-1919*, Second edn, Canberra, A. Kennedy.

Kopelman, N.M., Mayzel, J., Jakobsson, M., Rosenberg, N.A. & Mayrose, I. 2015. Clumpak: a program for identifying clustering modes and packaging population structure inferences across K, *Mol Ecol Resour*, 15, 1179-91.

Lao, O., Vallone, P.M., Coble, M.D., Diegoli, T.M., van Oven, M., van der Gaag, K.J., Pijpe, J., de Knijff, P. & Kayser, M. 2010. Evaluating Self-declared Ancestry of U.S. Americans with Autosomal, Y-chromosomal and Mitochondrial DNA, *Human Mutat*, 31, e1875-e93.

Lee, H.Y., Song, I., Ha, E., Cho, S.B., Yang, W.I. & Shin, K.J. 2008. mtDNAmanager: a Web-based tool for the management and quality analysis of mitochondrial DNA control-region sequences, *BMC Bioinformatics*, 9, 483.

Loogvali, E.L., Roostalu, U., Malyarchuk, B.A., Derenko, M.V., Kivisild, T., Metspalu, E., Tambets, K., Reidla, M., Tolk, H.V., Parik, J., et al. 2004. Disuniting uniformity: a pied cladistic canvas of mtDNA haplogroup H in Eurasia, *Mol Biol Evol*, 21, 2012-21.

Lutz, S., Weisser, H.J., Heizmann, J. & Pollak, S. 1998. Location and frequency of polymorphic positions in the mtDNA control region of individuals from Germany, *Int J Legal Med*, 111, 67-77.

Malyarchuk, B., Derenko, M., Grzybowski, T., Perkova, M., Rogalla, U., Vanecek, T. & Tsybovsky, I. 2010. The peopling of Europe from the mitochondrial haplogroup U5 perspective, *PLoS One*, 5, e10285.

Morales-Arce, A.Y., Hofman, C.A., Duggan, A.T., Benfer, A.K., Katzenberg, M.A., McCafferty, G. & Warinner, C. 2017. Successful reconstruction of whole mitochondrial genomes from ancient Central America and Mexico, *Sci Rep*, 7, 18100.

Musgrave-Brown, E., Ballard, D., Balogh, K., Bender, K., Berger, B., Bogus, M., Børsting, C., Brion, M., Fondevila, M., Harrison, C., et al. 2007. Forensic validation of the SNPforID 52-plex assay, *Forensic Sci Int Genet*, 1, 186-90.

Nagle, N., van Oven, M., Wilcox, S., van Holst Pellekaan, S., Tyler-Smith, C., Xue, Y., Ballantyne, K.N., Wilcox, L., Papac, L., Cooke, K., et al. 2017. Aboriginal Australian mitochondrial genome variation – an increased understanding of population antiquity and diversity, *Sci Rep*, 7, 43041.

Olivieri, A., Pala, M., Gandini, F., Kashani, B.H., Perego, U.A., Woodward, S.R., Grugni, V., Battaglia, V., Semino, O., Achilli, A., et al. 2013. Mitogenomes from Two Uncommon Haplogroups Mark Late Glacial/Postglacial Expansions from the Near East and Neolithic Dispersals within Europe, *PLoS One*, 8, e70492.

Pala, M., Olivieri, A., Achilli, A., Accetturo, M., Metspalu, E., Reidla, M., Tamm, E., Karmin, M., Reisberg, T., Hooshiar Kashani, B., et al. 2012. Mitochondrial DNA signals of late glacial recolonization of Europe from near eastern refugia, *Am J Hum Genet*, 90, 915-24.

Parkes, R. 2009. Traces of the cameleers: Landscape archaelogy and landscape perception, *Australasian Historical Archaeology*, 27, 87-97.

Parson, W. & Bandelt, H.J. 2007. Extended guidelines for mtDNA typing of population data in forensic science, *Forensic Sci Int Genet*, 1, 13-9.

Parson, W. & Dur, A. 2007. EMPOP--a forensic mtDNA database, *Forensic Sci Int Genet*, 1, 88-92.

Pereira, L., Richards, M., Goios, A., Alonso, A., Albarran, C., Garcia, O., Behar, D.M., Golge, M., Hatina, J., Al-Gazali, L., et al. 2005. High-resolution mtDNA evidence for the late-glacial resettlement of Europe from an Iberian refugium, *Genome Res*, 15, 19-24.

Phillips, C. 2015. Forensic genetic analysis of bio-geographical ancestry, *Forensic Sci Int Genet*, 18, 49-65.

Phillips, C., Prieto, L., Fondevila, M., Salas, A., Gómez-Tato, A., Álvarez-Dios, J., Alonso, A., Blanco-Verea, A., Brión, M., Montesino, M., et al. 2009. Ancestry Analysis in the 11-M Madrid Bomb Attack Investigation, *PLoS One*, 4, e6583.

Plaza, S., Calafell, F., Helal, A., Bouzerna, N., Lefranc, G., Bertranpetit, J. & Comas, D. 2003. Joining the Pillars of Hercules: mtDNA Sequences Show Multidirectional Gene Flow in the Western Mediterranean, *Ann Hum Genet*, 67, 312-28.

Porras-Hurtado, L., Ruiz, Y., Santos, C., Phillips, C., Carracedo, A. & Lareu, M.V. 2013. An overview of STRUCTURE: applications, parameter settings, and supporting software, *Front Genet*, 4, 98.

Prestes, P.R., Mitchell, R.J., Daniel, R., Sanchez, J.J. & van Oorschot, R.A.H. 2016. Predicting biogeographical ancestry in admixed individuals – values and limitations of using uniparental and autosomal markers, *Aust J Forensis Sci*, 48, 10-23.

Rando, J.C., Pinto, F., Gonzales, A.M., Hernandez, M., Larruga, J.M., Cabrera, V.M. & Bandelt, H.J. 1998. Mitochondrial DNA analysis of Northwest African populations reveals genetic exchanges with European, Near-Eastern, and sub-Saharan populations, *Ann Hum Genet*, 62, 531-50.

Richard, C., Pennarun, E., Kivisild, T., Tambets, K., Tolk, H.V., Metspalu, E., Reidla, M., Chevalier, S., Giraudet, S., Lauc, L.B., et al. 2007. An mtDNA perspective of French genetic variation, *Ann Hum Biol*, 34, 68-79.

Richards, M., Macaulay, V., Hickey, E., Vega, E., Sykes, B., Guida, V., Rengo, C., Sellitto, D., Cruciani, F., Kivisild, T., et al. 2000. Tracing European founder lineages in the Near Eastern mtDNA pool, *Am J Hum Genet*, 67, 1251-76.

Riseman, N. 2013. Serving their country: A short history of Aboriginal and Torres Strait Islander Service in the Australian Army, *Australian Army Journal*, 5, 11-22.

Roostalu, U., Kutuev, I., Loogvali, E.L., Metspalu, E., Tambets, K., Reidla, M., Khusnutdinova, E.K., Usanga, E., Kivisild, T. & Villems, R. 2007. Origin and expansion of haplogroup H, the dominant human mitochondrial DNA lineage in West Eurasia: the Near Eastern and Caucasian perspective, *Mol Biol Evol*, 24, 436-48.

Santos, C., Phillips, C., Fondevila, M., Daniel, R., van Oorschot, R.A.H., Burchard, E.G., Schanfield, M.S., Souto, L., Uacyisrael, J., Via, M., et al. 2016. Pacifiplex: an ancestry-informative SNP panel centred on Australia and the Pacific region, *Forensic Sci Int Genet*, 20, 71-80.

Steele, C.D. & Balding, D.J. 2014. Choice of population database for forensic DNA profile analysis, *Science & Justice*, 54, 487-93.

Taylor, D., Bright, J., McGovern, C., Neville, S. & Grover, D. 2017. Allele frequency database for GlobalFiler™ STR loci in Australian and New Zealand populations, *Forensic Sci Int Genet*, 28, e38-e40.

The 1000 Genomes Project Consortium. 2015. A global reference for human genetic variation, *Nature*, 526, 68.

Tobler, R., Rohrlach, A., Soubrier, J., Bover, P., Llamas, B., Tuke, J., Bean, N., Abdullah-Highfold, A., Agius, S., O'Donoghue, A., et al. 2017. Aboriginal mitogenomes reveal 50,000 years of regionalism in Australia, *Nature*, 544, 180.

Turchi, C., Buscemi, L., Previdere, C., Grignani, P., Brandstatter, A., Achilli, A., Parson, W. & Tagliabracci, A. 2008. Italian mitochondrial DNA database: results of a collaborative exercise and proficiency testing, *Int J Legal Med*, 122, 199-204.

Turchi, C., Stanciu, F., Paselli, G., Buscemi, L., Parson, W. & Tagliabracci, A. 2016. The mitochondrial DNA makeup of Romanians: A forensic mtDNA control region database and phylogenetic characterization, *Forensic Sci Int Genet*, 24, 136-42.

van Oven, M. 2015. PhyloTree Build 17: Growing the human mitochondrial DNA tree, *Forensic Sci Int Genet Supp Series*, 5, e392-e4.

van Oven, M., Ralf, A. & Kayser, M. 2011a. An efficient multiplex genotyping approach for detecting the major worldwide human Y-chromosome haplogroups, *Int J Legal Med*, 125, 879-85.

van Oven, M., Vermeulen, M. & Kayser, M. 2011b. Multiplex genotyping system for efficient inference of matrilineal genetic ancestry with continental resolution, *Invest Genet*, 2.

Vigilant, L., Stoneking, M., Harpending, H., Hawkes, K. & Wilson, A.C. 1991. African populations and the evolution of human mitochondrial DNA, *Science*, 253, 1503-7.

Walsh, P.S., Metzger, D.A. & Higuchi, R. 1991. Chelex 100 as a medium for simple extraction of DNA for PCR-based typing from forensic material, *Biotechniques*, 10, 506-13.

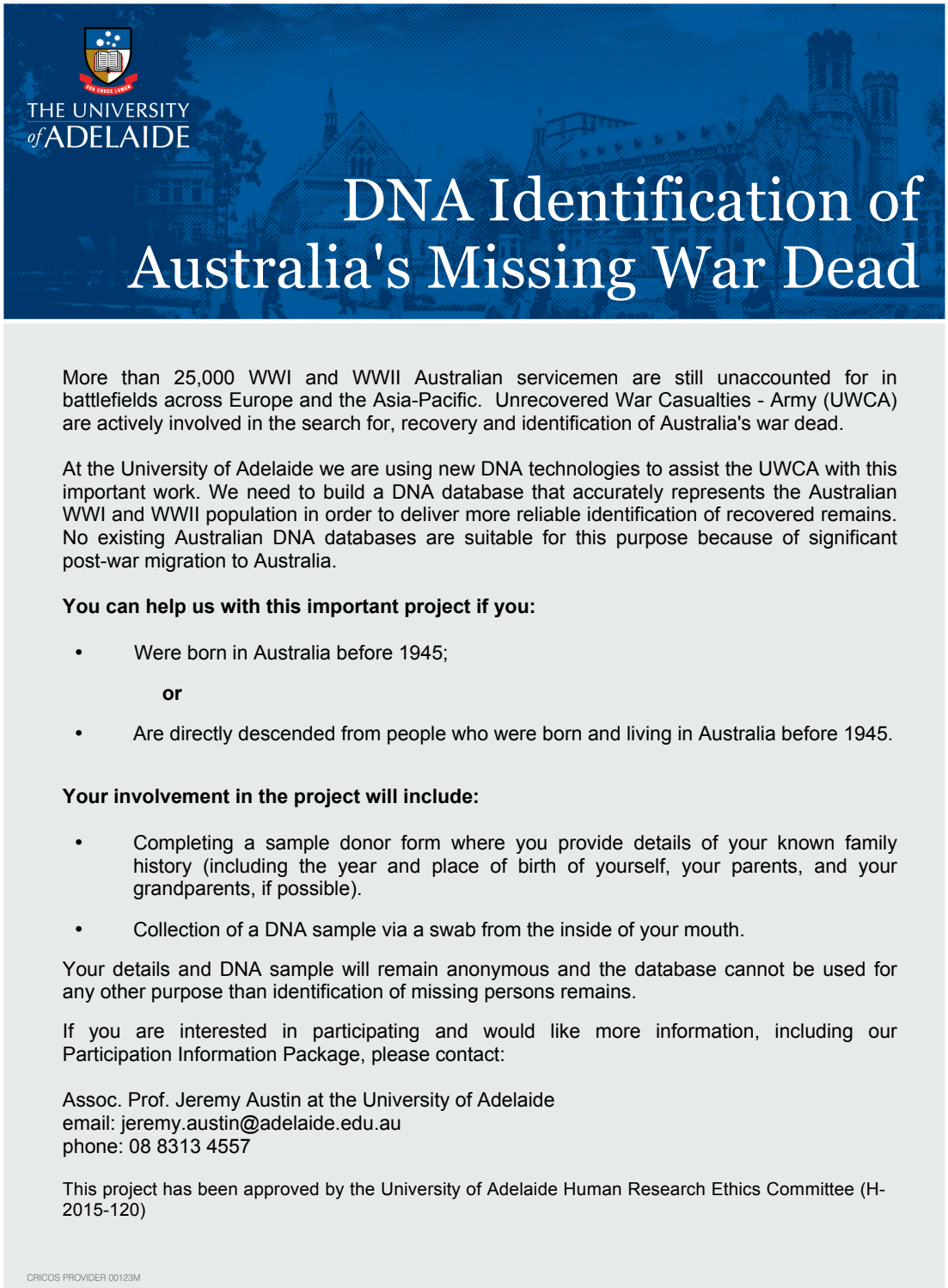
Weissensteiner, H., Pacher, D., Kloss-Brandstatter, A., Forer, L., Specht, G., Bandelt, H.J., Kronenberg, F., Salas, A. & Schonherr, S. 2016. HaploGrep 2: mitochondrial haplogroup classification in the era of high-throughput sequencing, *Nucleic Acids Res*, 44, W58-63.

Wilson, E.B. 1927. Probable Inference, the Law of Succession, and Statistical Inference, *J Am Stat Assoc*, 22, 209-12.

Yaacob, H., Nambiar, P. & Naidu, M.D. 1996. Racial characteristics of human teeth with special emphasis on the Mongoloid dentition, *Malays J Pathol*, 18, 1-7.

Supplementary Information

Supplementary File S1. Information flyer for public recruitment of suitable DNA donors for the Historical Australian DNA Database



Supplementary File S2. mtDNA control region data for 259 samples. Variant positions from the rCRS are shown between 16024 – 16569 and 1 – 576 (excluding length variations at 309, 515-522, 573 and 16193). Haplogroup nomenclature is according to PhyloTree Build 17.

Sample	Control-region variants (listed as differences to the rCRS)	Sub-haplogroup (PhyloTree Build 17)	Haplogroup	Broad Haplogroup
1	16311C 16362C 16482G 239C 263G	H6a1a1a	H6a	H
2	16093C 16224C 16234T 16278T 16311C 16320T 16519C 73G 263G 490T 497T	K1a	K1a	K
3	16126C 16163G 16186T 16189C 16294T 16519C 73G 152C 195C 263G	T1a1'3	T1a	T
4	16224C 16311C 16519C 73G 195C 234G 263G 497T	K1a	K1a	K
5	16298C 72C 263G	HV0	HV0	HV0
6	16126C 16163G 16186T 16189C 16294T 16519C 73G 195C 263G	T1a1'3	T1a	T
7	16069T 16126C 16145A 16172C 16192T 16222T 16261T 42.1T 73G 242T 263G 295T 462T 489C	J1b1a1a	J1b	J
8	16234T 16311C 150T 152C 263G	HV15	HV15	HV
9	16298C 72C 263G	HV0	HV0	HV0
10	16354T 16519C 152C 263G	H2a1	H2a	H
11	16172C 16192T 16456A 16519C 207A 263G	H1aj1	H1a	H
12	16354T 263G	H2a1	H2a	H
13	16126C 16182C 16183C 16189C 16294T 16296T 16298C 16519C 73G 195C 263G	T2f1a	T2f	T
14	16158G 16263C 16519C 263G 477C 513A	H1c1	H1c	H
15	16069T 16126C 73G 185A 188G 228A 263G 295T 462T 489C	J1c2	J1c	J
16	16184T 16223T 16292T 16519C 73G 119C 189G 195C 204C 207A 263G	W1c1	W1c	W
17	16192T 16270T 16304C 16526A 73G 150T 228A 263G	U5b3b	U5b	U
18	16224C 16311C 16362C 16519C 73G 195C 263G	K1d	K1d	K
19	16129A 16223T 16519C 73G 152C 204C 207A 250C 263G	I2	I2	I
20	16069T 16126C 16145A 16235G 16261T 16519C 73G 152C 207A 263G 271T 295T 462T 489C	J1b1b2	J1b	J
21	16129A 16519C 263G	H	H	H
22	16126C 16172C 16294T 16304C 16519C 73G 195C 263G 499A	T2b4a	T2b	T
23	16093C 16126C 16145A 16243C 16292T 16294T 16519C 73G 146C 152C 263G 279C	T2c1d	T2c	T
24	16183C 16189C 16223T 16278T 16292T 16519C 73G 143A 195C 198T 225A 226C 263G	X2b	X2b	X
25	16362C 16482G 239C 263G	H6a1	H6a	H
26	16069T 16126C 16145A 16172C 16189C 16192T 16222T 16261T 16311C 73G 242T 263G 295T 462T 489C	J1b1a1a	J1b	J
27	16126C 16298C 16346C 72C 200G 263G	H7a1	H7a	H
28	16311C 16519C 73G 263G 295A	R1a	R1a	R
29	16069T 16126C 16145A 16231C 16261T 73G 150T 152C 195C 215G 263G 295T 319C 489C 513A	J2a1a1a	J2a	J
30	**16093Y** 16224C 16311C 16519C 52C 73G 150T 195C 263G 497T	K1a	K1a	K

31	16293G 16519C 263G	H24	H24	H
32	16126C 16163G 16186T 16189C 16294T 16519C 73G 152C 183G 195C 263G	T1a1'3	T1a	T
33	16051G 16092C 16129C 16182C 16183C 16189C 16362C 16519C 73G 146C 152C 217C 263G 508G	U2e2	U2e	U
34	16304C 146C 152C 195C 249G 263G 456T	H5b1	H5b	H
35	263G	H	H	H
36	16298C 16311C 72C 263G	HV0	HV0	HV0
37	16224C 16245T 16311C 16519C 73G 146C 263G	K1a4a1a2b	K1a	K
38	16222T 16224C 16270T 16311C 16519C 73G 146C 263G	K2b1a1	K2b	K
39	16261T 16278T 16519C 152C 263G	H*	H	H
40	16356C 16519C 73G 152C 195C 263G 499A	U4b1b1b	U4b	U
41	16069T 16126C 16145A 16172C 16222T 16261T 73G 242T 263G 295T 462T 489C	J1b1a1	J1b	J
42	16069T 16126C 73G 185A 228A 263G 295T 462T 489C	J1c	J1c	J
43	73G 200G 263G	R	R	R
44	16067T 16183C 16189C 16519C 152C 263G	HV1b2	HV1	HV
45	16192T 16270T 16304C 73G 150T 263G	U5b3	U5b	U
46	16298C 263G	V	V	V
47	16162G 16519C 73G **195Y** 263G	H1a	H1a	H
48	16126C 16189C 16278T 16294T 16296T 16519C 73G 263G	T2f3	T2f	T
49	16239T 16519C 263G	H1	H1	H
50	16126C 16163G 16186T 16189C 16287T 16294T 16519C 73G 152C 195C 263G	T1a1'3	T1a	T
51	16183C 16189C 16356C 16362C 16519C 263G	H1b1	H1b	H
52	16224C 16311C 16320T 16519C 73G 146C 152C 263G 498d	K1c2	K1c	K
53	16114A 16192T 16256T 16270T 16294T 16526A 73G 152C 263G	U5a2a	U5a	U
54	16129A	H	H	H
55	16311C 195C 263G	H11	H11	H
56	16069T 16126C 16145A 16172C 16192T 16222T 16261T 73G 242T 263G 295T 462T 489C	J1b1a1a	J1b	J
57	16519C 93G 263G	H1e1a1	H1e	H
58	16179T 16354T 263G	H2a1	H2a	H
59	16069T 16126C 16167T 73G 184T 185A 228A 263G 295T 462T 489C	J1c	J1c	J
60	16069T 16126C 73G 185A 263G 295T 462T 489C	J1c	J1c	J
61	16192T 16256T 16269G 16270T 16291T 16294T 16399G 73G 263G	U5a1b1e	U5a	U
62	16093C 16299G 16519C 263G	H39	H39	H
63	**16093Y** 16311C 263G	H3h	H3h	H
64	16126C 16223T 16294T 16296T 16304C 16519C 73G 199C 204C 263G	T2b	T2b	T

65	16093C 16221T 16519C 263G	H10e	H10e	H
66	16069T 16126C 16519C 73G 185A 263G 295T 462T 489C	J1c	J1c	J
67	16069T 16126C 16145A 16172C 16192T 16222T 16261T 73G 242T 263G 295T 462T 489C	J1b1a1a	J1b	J
68	16356C 16519C 73G 195C 263G 499A	U4	U4	U
69	16162G 16209C 16519C 73G 263G	H1a1	H1a	H
70	16063C 16069T 16093C 16126C 73G 228A 263G 295T 462T 489C	J1c3f	J1c	J
71	16126C 16294T 16296T 16304C 16519C 73G 152C 263G	T2b	T2b	T
72	16519C 152C 263G	H	H	H
73	16189C 16342C 16519C 263G	H7d5	H7d	H
74	16519C 73G 195C 263G 310C 499A	U4a2a	U4a	U
75	16519C 263G 477C	H1c	H1c	H
76	16213A 16519C 263G	H7h	H7h	H
77	**16111Y** 16519C 152C 263G 477C	H1c	H1c	H
78	16187T 16222T 16224C 16270T 16311C 16519C 73G 146C 263G	K2b1a1	K2b	K
79	16129A 73G 195C 263G	H4/R8	H4	H
80	16224C 16311C 16362C 16474A 16519C 73G 146C 152C 189G 263G 498d	K1c	K1c	K
81	16192T 16256T 16270T 16526A 73G 263G	U5a2	U5a	U
82	16069T 16126C 73G 228A 263G 295T 462T 489C	J1c	J1c	J
83	16172C 16224C 16311C 16519C 73G 263G 497T	K1a	K1a	K
84	16311C 131C 152C 263G	HV9a	HV9	HV
85	16093C 16298C 72C 263G 513A	HV0	HV0	HV0
86	16111T 16519C 263G	H*	H	H
87	16298C 16526A 72C 263G	HV0	HV0	HV0
88	16069T 16126C 16311C 16519C 73G 185A 188G 228A 263G 295T 462T 489C	J1c2	J1c	J
89	16176T 16219G 16519C 146C 257G 263G 477C	H1c3a	H1c	H
90	16093C 16298C 72C 263G 513A	HV0	HV0	HV0
91	16304C 146C 195C 263G 456T	H5b1	H5b	H
92	16093C 16224C 16311C 16519C 73G 114T 263G 497T	K1a1	K1a	K
93	16126C 16294T 16296T 16304C 16519C 73G 263G	T2b	T2b	T
94	16069T 16126C 16145A 16235G 16261T 16519C 73G 152C 207A 263G 271T 295T 462T 489C	J1b1b2	J1b	J
95	16126C 16163G 16186T 16189C 16294T 16519C 73G 152C 263G	T1a	T1a	T
96	16126C 16257T 16270A 16294T 16296T 16519C 73G 151T 263G	T2	T2	T
97	16069T 16126C 73G 185A 228A 263G 295T 462T 489C	J1c	J1c	J
98	16298C 16301T 72C 263G	V16	V16	V

99	16292T 16519C 152C 263G	H	H*	V
100	16355T 16519C 150T 263G	H1o	H1o	H
101	16192T 16256T 16270T 16399G 73G 263G	U5a1	U5a	U
102	16355T 16519C 150T 263G	H1o	H1o	H
103	16126C 16239T 16294T 16304C 16519C 73G 152C 263G	T2b7a2	T2b	T
104	16217C 16218T 16519C 263G	H20	H20	H
105	16183C 16189C 16270T 16362C 16519C 73G 93G 150T 263G	U5b2b1a	U5b	U
106	16069T 16126C 16145A 16172C 16207G 16222T 16261T 64T 73G 150T 189G 242T 263G 295T 462T 489C	J1b1a1	J1b	J
107	16519C 263G 477C	H1c2	H1c	H
108	16162G 16209C 16519C 73G 146C 263G	H1a1	H1a	H
109	16093C 16519C 263G	H	H	H
110	16069T 16126C 73G 263G 295T 462T 489C	J1	J1	J
111	16214T 16343G 16362C 73G 150T 195C 263G 298T	U3b2	U3b	U
112	16298C 16325C 72C 195C 263G	HV0	HV0	HV0
113	16111T 16129A 16256T 16519C 153G 263G	H3b1b1	H3b	H
114	16069T 16126C 16234T 73G 185A 263G 295T 462T 489C	J1c	J1c	J
115	16519C 263G	H	H	H
116	263G	H	H	H
117	16189C 16192T 16249C 16270T 73G 150T 246C 263G	U5b2c2b	U5b	U
118	16354T 210G 263G	H2a1	H2a	H
119	16311C 152C 263G	H	H	H
120	16126C 16294T 16304C 16519C 73G 152C 263G	T2b	T2b	T
121	16069T 16126C 73G 185A 228A 263G 295T 462T 489C	J1c	J1c	J
122	16220C 16519C 263G	H1bs	H1b	J
123	16126C 16183C 16189C 16294T 16296T 16298C 16519C 73G 195C 263G	T2f1a	T2f	T
124	16304C 263G 456T	H5	H5	H
125	16126C 16222A 16294T 16296T 16304C 16519C 73G 263G	T2b	T2b	T
126	16519C 263G	H	H	H
127	16189C 16270T 16311C 73G 150T 263G	U5b2a5	U5b	U
128	16519C 146C 263G	H	H	H
129	16126C 16172C 16254G 16294T 16304C 16519C 73G 263G	T2b4a	T2b	T
130	16093C 16362C 16519C 152C 263G 408A	H3v + 16093	H3v	H
131	16291T 16343G 16390A 16519C 207A 263G	H85	H85	H
132	16051G 16092C 16129C 16182C 16183C 16189C 16362C 16519C 73G 152C 195C 217C 263G 508G	U2e2	U2e	U

133	263G	H	H	H
134	16223T 16292T 16362C 16519C 73G 152C 189G 194T 195C 204C 207A 263G 496T	W5a1a	W5a	W
135	16223T 16292T 16325C 16519C 73G 189G 194T 195C 204C 207A 263G 552T	W6	W6	W
136	16519C 263G	H	H	H
137	16298C 72C 263G	HV0	HV0	HV0
138	16126C 16189C 16278T 16294T 16296T 16519C 73G 263G	T2f3	T2f	T
139	16069T 16126C 16145A 16172C 16192T 16222T 16261T 73G 242T 263G 295T 462T 489C	J1b1a1a	J1b	J
140	16304C 16519C 263G 456T	H5	H5	H
141	16069T 16126C 73G 263G 295T 462T 489C	J1	J1	J
142	16209C 16304C 16519C 263G 456T	H5a1j	H5a	H
143	16519C 195C 263G	H1bi	H1b	H
144	16224C 16311C 16519C 73G 146C 152C 263G 498d	K1c	K1c	K
145	16129A 16519C 263G	H	H	H
146	16162G 16209C 16519C 73G 263G	H1a1	H1a	H
147	16069T 16126C 73G 185A 263G 295T 462T 489C	J1c	J1c	J
148	16126C 16183C 16189C 16294T 16296T 16519C 73G 152C 263G	T2	T2	T
149	16189C 16519C 263G	H1	H1	H
150	16265G 16356C 16362C 16519C 73G 195C 247A 263G 499A	U4a3	U4a	U
151	16261T 16356C 16519C 73G 195C 263G 310C 499A	U4a2	U4a	U
152	16126C 16294T 16304C 16519C 73G 199C 263G 321C	T2b24	T2b	T
153	16183C 16189C 16223T 16278T 16519C 73G 153G 195C 225A 226C 263G 447T	X2b	X2b	X
154	16119G 16224C 16311C 16519C 73G 263G 497T	K1a	K1a	K
155	16162G 16519C 73G 263G 513A	H1a	H1a	H
156	16304C 93G 263G 456T	H5	H5	H
157	16075C 16117C 16129A 16209C 16223T 16318T 16519C 73G 150T 152C 252C 263G	S1	S1	S
158	16235G 16291T 16519C 263G	H2a2b	H2a	H
159	16145A 16224C 16311C 16519C 73G 263G 497T	K1a	K1a	K
160	16189C 16356C 16362C 16519C 263G	H1b1	H1b	H
161	16126C 16158G 16163G 16186T 16189C 16294T 16519C 73G 152C 195C 263G	T1a	T1a	T
162	16291T 16519C 263G	H2a5a1	H2a	H
163	16093C 16519C 263G 408A	H3v	H3v	H
164	16224C 16311C 16519C 73G 146C 195C 263G	K1b2	K1b	K
165	16093C 16224C 16256T 16311C 16319A 16463G 16519C 73G 152C 263G	K1b1a	K1b	K
166	16223T 16292T 16519C 73G 189G 195C 204C 207A 263G	W	W	W

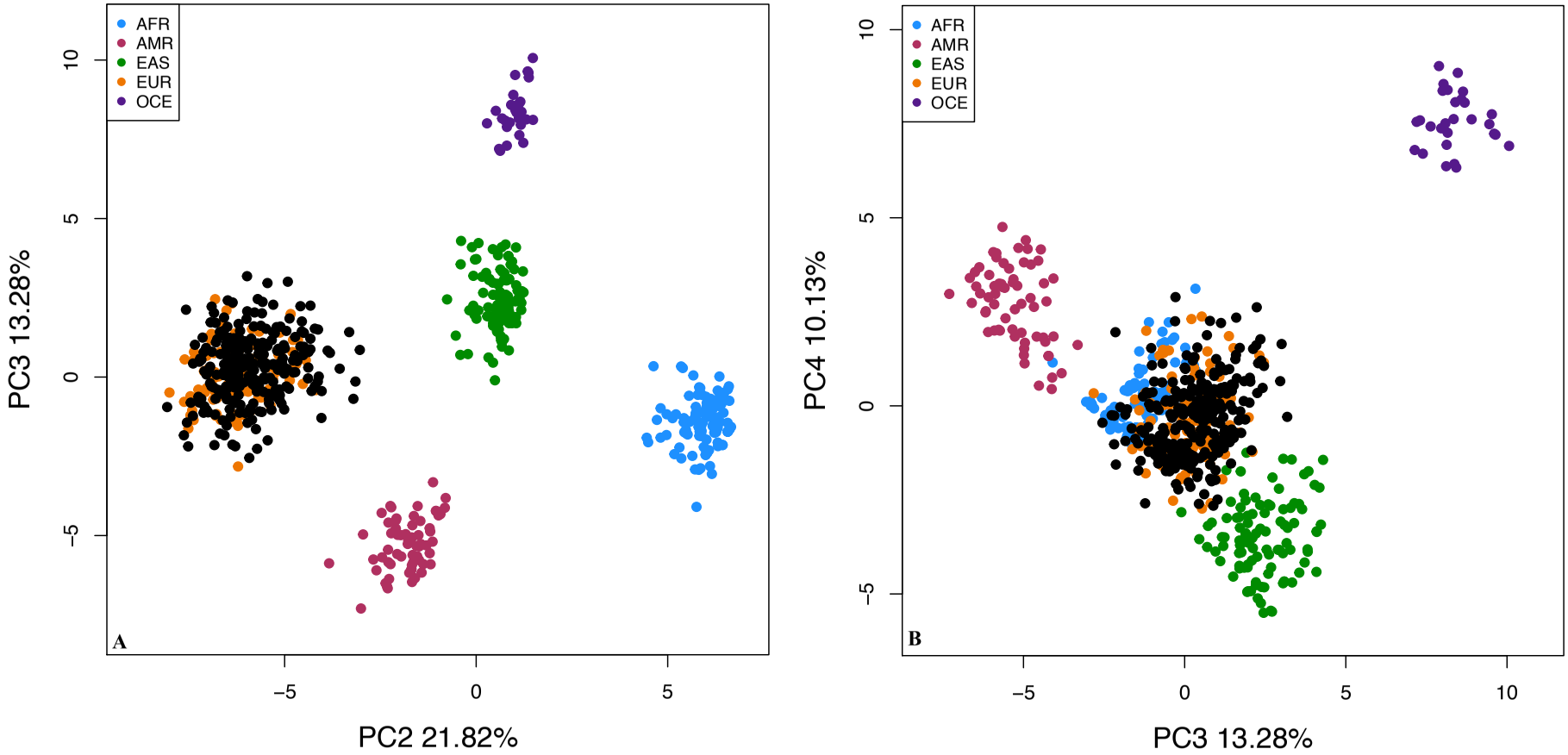
167	16192T 16256T 16270T 16311C 16399G 16526A 73G 199C 263G	U5a1f1a1	U5a	U
168	16311C 195C 263G	H11	H11	H
169	16519C 252C 263G	H	H	H
170	16126C 16256T 16294T 16519C 73G 263G	T	T	T
171	16069T 16126C 16209C 16265G 16319A 73G 185A 189G 263G 295T 462T 489C	J1c8a1	J1c	J
172	16519C 263G	H	H	H
173	263G 383C	H	H	H
174	16037G 16126C 16153A 16245A 16294T 16296T 16298C 16519C 73G 150T 263G	T2e	T2e	T
175	16189C 16311C 16354T 195C 263G	HV8	HV8	HV
176	16148T 16519C 263G	H1au	H1a	H
177	16126C 16292T 16294T 16296T 16519C 73G 146C 263G	T2c1	T2c	T
178	16140C 16293G 16311C 195C 263G	H11a2a	H11a	H
179	16126C 16235G 16291T 263G	H2a2b	H2a	H
180	16291T 16311C 263G	H2a	H2a	H
181	16256T 16270T 16399G 73G 263G	U5a1	U5a	U
182	16192T 16256T 16270T 16320T 16399G 73G 195C 263G	U5a1c1	U5a	U
183	16129A 16172C 16223T 16311C 16391A 16519C 73G 199C 202G 203A 204C 250C 263G 455.1T 574C	I1a1	I1a	I
184	16162G 16298C 16519C 72C 263G	V6	V6	V
185	16224C 16311C 16519C 73G 263G 497T	K1a	K1a	K
186	16162G 16189C 16266T 16519C 73G 151T 263G	H1a6	H1a	H
187	16172C 16183C 16189C 16219G 16239T 16278T 16362C 73G 263G	U6a1a1	U6a	U
188	16298C 72C 263G	HV0	HV0	HV0
189	16136C 16295T 16519C 263G	H16b	H16b	H
190	16519C 93G 263G 477C	H1c	H1c	H
191	152C 263G	H*	H	H
192	16189C 16356C 16519C 263G 408A	H1b	H1b	H
193	16051G 16519C 152C 263G	H1i	H1i	H
194	16192T 16256T 16270T 16290T 16291T 16399G 73G 263G 466C	U5a1b1e	U5a	U
195	16126C 16163G 16186T 16189C 16294T 16519C 73G 111G 152C 195C 263G	T1a1'3	T1a	T
196	16037G 16093C 16519C 263G	H1q3	H1q	H
197	16111T 16129A 16172C 16223T 16311C 16391A 16519C 73G 199C 203A 204C 250C 263G 455.1T 574C	I1a1	I1a	I
198	16304C 263G 456T	H5	H5	H
199	16182C 16183C 16189C 16519C 263G	H1	H1	H
200	16189C **16192Y** 16270T 16398A 73G 150T 263G	U5b2a2	U5b	U

201	16114T 16278T 16519C **215R** 263G	H1e4a	H1e	H
202	16069T 16126C 16153A 16390A 73G 263G 295T 462T 489C	J1	J1	J
203	16189C 16207G 16356C 16497G 16519C 263G	H1b	H1b	H
204	16519C 16527T 195C 257G 263G 477C	H1c3	H1c	H
205	16184A 16224C 16311C 16519C 73G 114T 146C **189R** 263G 497T	K1a1b1	K1a	K
206	16256T 16270T 16399G 73G 263G 574C	U5a1a2	U5a	U
207	16126C 16294T 16304C 16519C 73G 263G	T2b	T2b	T
208	16069T 16126C 16145A 16172C 16222T 16261T 73G 185A 242T 263G 295T 462T 489C	J1b1a1d	J1b	J
209	16519C 152C 263G	H	H	H
210	16086C 16147A 16223T 16248T 16320T 16355T 16519C 73G 152C 199C 204C 207A 263G 547T	N1a1a1a2	N1a	N
211	16213A 16519C 263G	H7h	H7h	H
212	16114A 16192T 16256T 16270T 16294T 16526A 73G 152C 263G	U5a2a	U5a	U
213	16192T 16249C 16270T 16311C 73G 150T 263G	U5b2c2b	U5b	U
214	16291T 16519C 73G 263G 453C	H1e1b	H1e	H
215	16519C 204C 263G	H	H	H
216	16069T 16126C 73G 185A 228A 263G 295T 462T 482C 489C	J1c1	J1c	J
217	16051G 16162G 16213A 16266T 16519C 73G 263G	H1a3c1	H1a	H
218	16189C 16356C 16362C 16519C 263G	H1b1	H1b	H
219	16235G 16291T 16292T 16319A 263G	H2a2b	H2a	H
220	16069T 16126C 16145A 16172C 16222T 16261T 73G 242T 263G 295T 462T 489C	J1b1a1	J1b	J
221	16519C 263G 385G	HV	HV	HV
222	16223T 16519C 152C 263G	R	R	R*
223	16167T 16179T 73G 263G 282C	U8a1b	U8a	U
224	16519C 152C 263G	H	H	H
225	16172C 16219G 16274A 16278T 73G 152C 263G	U6a7a	U6a	U
226	16183C 16189C 16218T 16311C 16519C 263G	H	H	H
227	16192T 16256T 16270T 16291T 16399G 73G 263G 466C	U5a1b1	U5a	U
228	16171T 16263C 16519C 263G 477C	H1c1	H1c	H
229	16223T 16292T 16519C 73G 189G 195C 204C 207A 263G	W	W	W
230	16519C 263G	H	H	H
231	16256T 16270T 16526A 73G 263G	U5a2c3	U5a	U
232	16069T **16092Y** 16126C 16519C 73G 185A 188G 228A 263G 295T 462T 489C	J1c2	J1c	J
233	16519C 152C 263G	H	H	H
234	16291T 16519C 263G	H2a5a1	H2a	H

235	16519C 150T 152C 263G	H	H	H
236	16192T 16256T 16270T 16320T 16399G 16527T 73G 153G 195C 263G	U5a1c1a	U5a	U
237	16311C 16362C 16482G 239C 263G	H6a1a1a	H6a	H
238	16183C 16189C 16217C 16325C 16519C 73G 263G 498d 499A	B2d	B2d	B
239	16519C 263G 499A	R	R	R*
240	16298C 72G 204C 263G	HV0	HV0	HV0
241	16093C 16224C 16311C 16327T 16519C 16T 73G 150T 200G 263G 497T	K1a	K1a	K
242	16224C 16311C 16519C 73G 146C 152C **195Y** 263G	K2a	K2a	K
243	16304C 263G 456T	H5	H5	H
244	263G	H	H	H
245	16192T 16256T 16270T 16320T 16399G 73G 183G 184A 215G 263G	U5a1c2a	U5a	U
246	16069T 16145A 16231C 16261T 73G 150T 152C 195C 215G 263G 295T 319C 489C 513A	J2a1a1	J2a	J
247	16184T 16223T 16292T 16519C 73G 189G 195C 204C 207A 225A 263G	W	W	W
248	16192T 16311C 73G 150T 263G	U5b2a1a	U5b	U
249	16519C 152C 263G	H	H	H
250	16069T 16126C 16362C 73G 89C 185A 195C 228A 263G 295T 462T 489C	J1c	J1c	J
251	16069T 16126C 73G 185A 263G 295T 408A 462T 489C	J1c	J1c	J
252	16093C 16519C 152C 263G	H	H	H
253	16126C **16188Y 16227R** 16257T 16294T 16296T 16519C 73G 235G 263G 507C	T2g2a	T2g	T
254	16354T 263G	H2a1	H2a	H
255	16519C 263G 293C	H3z	H3z	H
256	16188G 16519C 263G	H1q	H1q	H
257	16189C 16192T 16256T 16270T 16362C 16399G 16428A 16430G 73G 263G	U5a1b3	U5a	U
258	16224C 16311C 16519C 73G 146C 195C 263G	K1b2	K1b	K
259	16223T 16270T 16295T 16356C 16519C 73G 195C	U4a2	U4a	U

Supplementary File S3. Details of the 402 reference samples from five population groups across the 1000 Genomes and HGDP-CEPH datasets used for comparison. Genotypes are provided in the strand direction of primers used in the Global AIMs Nano set for direct comparison (provided as an electronic copy on USB Drive).

Supplementary File S4. (A) PC2 versus PC3, and (B) PC3 versus PC4 of database samples (black) against 402 genotypes across five global reference population groups.



Supplementary File S5. Average population membership proportions from STRUCTURE analysis of database samples (K = 5).

Sample	% AFR	% AMR	% EAS	% EUR	% OCE
1	0.24	0.2	0.26	99.06	0.22
2	5.42	0.2	0.28	93.88	0.26
3	0.22	0.1	0.2	99.28	0.24
4	0.34	0.26	0.3	98.94	0.13
5	0.26	0.1	0.2	99.3	0.16
6	1.46	0.44	0.66	96.76	0.66
7	0.2	0.26	0.2	99.14	0.22
8	0.28	0.94	0.46	98.08	0.24
9	0.18	0.1	0.22	99.34	0.1
10	0.2	0.3	0.38	99.02	0.1
11	0.24	0.2	0.48	98.78	0.32
12	0.3	0.1	0.2	99.24	0.2
13	0.4	0.48	0.28	98.38	0.46
14	0.2	0.1	0.18	99.36	0.2
15	0.38	0.1	0.18	99.2	0.12
16	17.24	0.22	0.56	80.76	1.2
17	0.4	0.98	0.52	97.6	0.46
18	0.32	0.1	0.1	99.28	0.18
19	0.2	0.22	0.58	98.86	0.14
20	0.2	0.1	0.2	99.22	0.24
21	0.44	0.1	0.2	99.04	0.2
22	2.56	0.1	0.4	96.14	0.84
23	0.3	0.24	0.28	98.92	0.26
24	0.2	0.18	0.22	99.28	0.12
25	8.8	0.88	1.26	88.62	0.42
26	0.46	1.6	1.4	95.86	0.6
27	0.68	0.14	0.2	98.7	0.26
28	9.84	0.42	0.5	88.7	0.54
29	1.4	0.54	1.9	95.3	0.88
30	0.2	0.1	0.12	99.42	0.1
31	0.2	0.1	0.16	99.36	0.12
32	4.06	1	17.64	71.02	6.28
33	0.46	0.14	0.3	98.62	0.46
34	0.38	0.2	0.2	99.08	0.12
35	0.28	0.14	0.34	99.1	0.2
36	0.32	0.1	0.2	99.28	0.1
37	0.2	0.1	0.22	99.36	0.14
38	0.3	0.3	0.66	98.26	0.52
39	0.32	0.18	0.6	98.76	0.18
40	1.4	1.46	2.9	93.5	0.74

Sample	% AFR	% AMR	% EAS	% EUR	% OCE
131	0.22	0.1	0.1	99.5	0.1
132	0.22	0.1	0.14	99.42	0.1
133	0.62	0.3	3.18	94.88	0.98
134	0.66	0.3	0.78	97.96	0.3
135	0.3	0.36	0.86	97.88	0.62
136	0.26	2.32	2.86	94.04	0.5
137	0.38	0.2	0.86	98.26	0.3
138	0.28	0.16	0.2	99.22	0.1
139	0.22	0.22	0.36	98.7	0.5
140	0.18	0.18	0.18	99.3	0.2
141	5.66	0.14	0.74	93.08	0.38
142	0.2	0.2	0.34	99.14	0.14
143	0.32	0.46	0.4	98.28	0.52
144	2.04	0.54	0.56	95.58	1.26
145	0.2	0.2	0.52	97.86	1.2
146	0.36	0.2	0.3	98.88	0.28
147	0.2	0.1	0.1	99.5	0.1
148	0.24	0.1	0.2	99.34	0.18
149	0.36	0.1	0.2	98.82	0.5
150	0.2	0.34	1.02	98.18	0.3
151	0.38	0.2	0.26	98.66	0.54
152	6.2	0.2	0.22	93.26	0.12
153	10.18	0.4	0.3	88.94	0.24
154	6.24	0.1	0.2	92.88	0.54
155	6.82	0.18	0.92	91.82	0.26
156	0.44	1.26	3.48	93.94	0.88
157	1.54	3.38	1.6	89.84	3.64
158	0.38	0.18	0.32	98.78	0.34
159	0.22	0.1	0.18	99.3	0.2
160	2.08	0.7	1.14	94.8	1.28
161	7.36	0.14	0.4	91.86	0.2
162	2.72	0.22	0.4	89.58	7.12
163	0.24	0.1	0.12	99.24	0.28
164	2.46	1.42	2.14	92.64	1.34
165	0.46	0.1	0.2	98.86	0.34
166	7.68	0.9	0.8	88.34	2.28
167	0.34	0.26	0.5	98.56	0.4
168	0.3	0.28	0.22	99.04	0.2
169	3	0.12	0.58	95.6	0.68
170	2.1	1.84	1.68	92.96	1.44

41	0.22	0.14	0.16	99.26	0.2
42	0.3	0.1	0.26	99.22	0.1
43	0.3	0.78	0.54	97.48	0.92
44	0.72	0.72	0.58	95.26	2.72
45	0.26	0.1	0.22	99.22	0.2
46	1.3	0.42	0.5	97.16	0.62
47	0.24	0.1	0.2	99.32	0.1
48	0.2	0.32	0.34	98.88	0.24
49	0.34	0.2	0.2	98.36	0.88
50	0.2	0.1	0.18	99.36	0.1
51	0.56	1.2	8.32	87.94	1.98
52	0.34	0.2	0.36	97.36	1.76
53	0.46	0.2	0.54	98.1	0.72
54	0.18	0.24	0.38	98.88	0.36
55	0.2	0.1	0.1	99.54	0.1
56	0.24	5.54	2.38	91.68	0.14
57	0.24	0.1	0.12	99.38	0.12
58	1.66	0.48	0.68	96.72	0.48
59	0.38	0.14	0.2	99.04	0.22
60	0.94	0.66	1.1	93.48	3.82
61	0.26	0.1	0.16	99.34	0.16
62	0.12	0.34	0.22	99.14	0.18
63	1.64	0.72	1.06	96.24	0.38
64	0.3	0.22	0.2	98.92	0.34
65	0.66	0.28	0.42	98.12	0.52
66	0.2	0.1	0.16	99.44	0.1
67	0.54	0.1	0.16	99.02	0.16
68	0.3	0.2	0.18	99.22	0.12
69	0.34	0.6	0.86	97.3	0.9
70	0.22	0.2	0.3	99.02	0.22
71	0.58	0.12	0.5	98.48	0.32
72	0.2	0.2	0.26	99.04	0.3
73	0.28	0.1	0.16	99.18	0.3
74	0.2	1.06	0.26	98.24	0.22
75	0.52	0.48	0.38	98.2	0.42
76	0.2	0.14	0.18	99.34	0.12
77	0.4	0.22	0.34	97.4	1.62
78	0.26	0.12	0.16	99.02	0.4
79	0.44	1.18	0.62	97.6	0.18
80	0.66	0.32	0.34	97.8	0.86
81	0.28	0.16	0.22	99.16	0.18
82	0.32	0.1	0.2	98.56	0.82
83	0.3	0.12	0.22	98.94	0.38
171	0.2	1.16	0.24	97.92	0.5
172	0.9	0.58	0.84	97	0.7
173	0.24	0.24	0.16	99.24	0.1
174	0.48	0.2	0.32	98.82	0.2
175	0.2	0.32	0.2	99.08	0.16
176	0.16	0.26	0.22	99.2	0.16
177	7.72	0.2	0.2	91.76	0.1
178	0.16	1.26	1.86	95.62	1.08
179	2.08	1.14	1.14	94.5	1.14
180	0.22	1.44	1.16	96.4	0.78
181	0.2	0.18	0.2	99.24	0.18
182	0.24	0.12	0.1	99.38	0.1
183	0.18	0.54	1.44	97.58	0.24
184	6.86	0.28	0.24	92.32	0.28
185	0.34	1.66	2.7	94.76	0.56
186	0.24	0.58	0.94	97.38	0.84
187	0.3	0.52	1.12	97.56	0.52
188	0.46	1.02	1.62	95.64	1.24
189	0.86	0.64	0.78	96.14	1.58
190	0.2	0.2	0.26	99.26	0.1
191	0.26	3.22	5.88	90.48	0.12
192	0.2	0.2	0.22	99.26	0.1
193	0.4	0.72	0.32	98.3	0.24
194	0.32	0.18	0.1	99.2	0.18
195	1.1	5.46	2.98	87.9	2.58
196	0.36	0.12	0.2	99.04	0.26
197	0.68	0.16	0.22	98.66	0.22
198	0.2	0.4	1.78	95.94	1.66
199	0.8	0.76	1.38	96.66	0.4
200	0.42	0.18	0.3	98.9	0.22
201	0.38	0.32	0.92	98.14	0.28
202	0.26	0.18	0.32	99.12	0.12
203	0.8	0.58	11.84	85.9	0.78
204	0.2	0.3	0.24	99.12	0.2
205	0.36	0.14	0.28	98.84	0.34
206	0.64	0.18	0.46	98.36	0.38
207	0.2	0.1	0.2	99.3	0.2
208	0.26	0.24	0.52	98.66	0.3
209	1.26	0.28	0.46	96.94	1.04
210	0.28	1.56	1.92	95.66	0.58
211	0.3	0.14	0.2	99.18	0.16
212	0.6	0.66	0.62	95.48	2.64
213	0.3	0.12	0.3	99.1	0.18

84	0.2	0.1	0.2	99.4	0.1
85	0.3	0.16	0.18	98.08	1.3
86	0.54	3.14	6.32	78.1	11.88
87	0.28	0.1	0.2	99.26	0.14
88	0.3	0.12	0.26	97.66	1.64
89	0.44	0.18	0.2	98.88	0.34
90	0.2	0.1	0.1	99.5	0.1
91	0.2	0.1	0.16	99.34	0.18
92	0.26	0.68	0.96	97.6	0.52
93	0.28	0.54	1.22	97.1	0.9
94	0.4	0.2	0.36	98.4	0.64
95	0.2	0.26	0.38	99	0.14
96	0.22	0.18	0.26	99.08	0.24
97	0.22	0.6	0.5	98.48	0.14
98	0.22	0.2	0.2	99.12	0.26
99	0.28	0.1	0.16	99.36	0.1
100	1.38	0.38	0.5	97.46	0.26
101	0.36	0.1	0.1	99.24	0.2
102	0.5	0.18	0.24	98.9	0.18
103	0.36	0.14	0.24	99.08	0.18
104	0.3	0.1	0.2	99.22	0.16
105	0.28	0.38	0.42	98.5	0.44
106	0.22	0.12	0.14	99.34	0.16
107	0.5	0.1	0.26	98.08	1.04
108	0.4	0.24	0.48	98.52	0.34
109	0.24	0.18	0.16	99.22	0.2
110	0.5	0.16	0.2	98.94	0.2
111	0.34	0.2	0.32	98.88	0.24
112	0.46	0.1	0.2	98.9	0.34
113	0.24	0.1	0.14	99.34	0.14
114	0.2	0.14	0.16	99.4	0.12
115	0.2	0.2	0.22	99.12	0.26
116	0.2	0.2	0.1	99.42	0.1
117	0.2	0.3	0.44	98.14	0.9
118	0.22	0.1	0.12	99.18	0.36
119	0.66	0.28	0.36	97.98	0.74
120	0.36	0.1	0.2	98.72	0.62
121	0.42	0.24	0.32	98.82	0.26
122	5.32	0.24	0.72	92.24	1.46
123	0.18	0.1	0.1	99.46	0.1
124	0.28	1.2	3.42	94.84	0.28
125	0.2	0.1	0.1	99.46	0.1
126	0.28	0.2	0.24	98.98	0.3
214	0.2	0.9	0.32	98.44	0.16
215	4.74	0.28	0.26	94.5	0.2
216	1.24	0.94	1.48	92.82	3.52
217	0.2	0.24	0.38	98.92	0.26
218	0.2	0.18	0.22	99.2	0.22
219	0.5	0.18	0.16	99.04	0.14
220	0.2	0.2	0.16	99.26	0.14
221	11.72	0.26	0.86	86.14	1
222	0.22	0.18	0.24	99.06	0.32
223	1.04	0.38	8.2	85.86	4.5
224	0.6	5.48	0.82	91.68	1.42
225	0.32	0.62	1.28	97.38	0.4
226	1.1	0.42	0.7	97.02	0.74
227	0.4	0.22	0.2	98.96	0.22
228	0.32	0.16	0.3	99.12	0.12
229	0.24	0.18	0.24	99.12	0.24
230	0.32	0.3	0.24	98.64	0.48
231	7.62	0.3	0.54	88.26	3.32
232	1.36	0.66	1.38	95.84	0.8
233	0.28	0.54	0.6	97.86	0.74
234	0.2	0.12	0.14	99.42	0.1
235	0.22	0.1	0.12	99.38	0.2
236	0.28	0.38	0.46	98.72	0.22
237	0.3	0.24	0.32	98.54	0.58
238	0.24	0.72	0.38	98.12	0.54
239	0.32	0.22	0.46	98.8	0.2
240	3.18	0.1	0.38	95.76	0.52
241	0.2	0.22	0.22	99.16	0.12
242	3.8	0.18	0.64	94.72	0.68
243	0.32	0.2	0.3	99.02	0.18
244	0.34	0.54	0.5	98.42	0.2
245	0.42	0.1	0.2	99.12	0.2
246	0.28	0.12	0.2	99.28	0.18
247	6.02	0.22	0.64	92.86	0.28
248	0.28	0.26	0.46	97.7	1.3
249	0.28	0.2	0.26	99.1	0.2
250	0.28	0.2	0.2	99.18	0.18
251	0.54	1.16	0.64	96.38	1.24
252	0.52	0.42	0.34	98.54	0.2
253	1.08	1.76	1.32	94.16	1.62
254	0.32	0.1	0.66	98.64	0.3
255	2.3	0.32	1.96	91.54	3.92
256	0.22	0.14	0.22	99.22	0.18

127	0.5	0.2	0.28	98.82	0.24
128	0.38	0.1	0.3	98.76	0.44
129	0.22	0.1	0.1	99.38	0.2
130	0.2	0.14	0.18	99.32	0.18

257	3.94	0.12	0.56	93.24	2.16
258	0.58	0.16	0.2	98.14	0.9
259	0.26	0.12	0.16	99	0.48

Chapter 6

General Discussion and Conclusion

Introduction and Thesis Summary

Identification of human remains is a complex task that is important for social, humanitarian and legal reasons. Conventional avenues for identification (both non-DNA and DNA-based) are not always successful for degraded or fragmentary remains, prompting the research and application of advanced genetic techniques to retrieve post-mortem information that can aid in forensic investigations of identity. Areas currently under intense research are the development of techniques that improve the genotyping success of the most challenging samples, where DNA is degraded and present in very low amounts, and the exploration of alternative sources of genetic intelligence data that may be of investigative value. These include the genotyping of SNPs for the prediction of biogeographic ancestry, and those that can predict hair and eye colour, both of which have previously demonstrated probative value for forensic investigation (Phillips *et al.* 2009). Genotyping advancements such as the advent of massively parallel sequencing has increased our ability to gain information for hundreds of forensically relevant SNPs including those for ancestry and phenotype, but still present issues in the analysis of highly degraded and compromised remains (Gettings *et al.* 2015; Elwick *et al.* 2018). Overall, such techniques have thus far been limited in their application to forensic casework in Australia involving degraded/fragmented remains from historical burials, cold cases and unidentified war dead recovered from past battlefields across World War I and II. The question of how Australia's migration history may impact on ancestry analysis is also important to address.

In an attempt to address these key aspects, the research reported in this thesis aimed to improve our current understanding of forensic intelligence testing of human remains in an Australian context. Furthermore, I explored and evaluated new techniques that can increase the amount of genetic information we can retrieve from forensically challenging samples to assist in investigations of identity. This research has investigated many aspects of ancestry testing, including the development, evaluation and application of novel genotyping workflows, and the investigation of genetic ancestry components in an historical Australian population. Overall, this thesis sought insight into three key areas:

1. To develop a simple, sensitive SNP typing tool to screen samples for DNA quality and broad biological profile before making decisions on downstream processing.

2. To develop, evaluate and apply a novel SNP typing panel using highly specialised DNA analysis technologies for improved genetic analysis of forensic intelligence for highly degraded human remains.
3. To estimate the biogeographic ancestry components in Australia before the major waves of migration in 1945, how the ancestry classifications differ between different marker sets, and to collate the first historical DNA database for Australia that will assist with war dead identification.

Summary

In Chapter 2 I assembled and developed a novel laboratory tool for the triaging of degraded human DNA based on DNA sample quality and broad biological profile. Eighteen markers were selected each for their informativeness for mtDNA and Y chromosome haplogroup, autosomal ancestry, and eye colour. This selection of markers was then developed into a complete laboratory workflow from amplification to data analysis. I showed that the panel created was able to successfully infer a range of broad ancestries, sex and eye colour and to successfully retrieve SNP data from a range of degraded human teeth samples. I demonstrated the value of the panel as an initial screening tool to either triage samples based on sample quality to avoid unnecessary sample consumption and re-analysis. The Miniplex has the ability to assist in deciding what downstream processes may be likely to retrieve sufficient genetic data, and to potentially streamline workflows by allowing the prioritisation of samples from further laborious laboratory processing.

Chapter 3 describes the development of a novel target enrichment panel for the retrieval of autosomal and Y chromosome SNP data from highly degraded DNA. The overall motivation for the development of this methodology is that it has the ability to retrieve very small DNA fragments that are not amendable to PCR-based enrichment methods. I evaluate this new panel for population differentiation of reference population groups with low admixture before applying it to a set of modern samples where biogeographic ancestry, sex, hair and eye colour were known. The study showed high accuracy for ancestry predictions with the exception of American and North African ancestry where modern populations have been previously characterised as admixed. However, the Y-chromosome SNPs in the panel aided in resolving the ancestry for these samples and further advocated for the use of multi-gene targets for the inference of ancestry. Hair and eye colour predictions were shown to align with error rates established by previous studies. Overall, the analysis showed that the genetic information

retrieved and the inferences made from this panel were robust and accurate for application to degraded casework DNA samples.

In Chapter 4 I explore the application of both the Miniplex and the custom hybridisation enrichment panel to a set of degraded human teeth and forensic casework samples. The Miniplex was shown to provide broad indications of biological profile, and to serve as a predictor of enrichment/MPS success. Both techniques were shown to retrieve genetic information and make congruent inferences of ancestry, sex, broad mtDNA lineage and phenotype from these samples, demonstrating the value of this novel approach in gaining intelligence data for forensic purposes.

In Chapter 5 I described the creation of the first forensic historical DNA database for the Australian population pre-1945. The information collected allows the characterisation of the genetic ancestry components that existed in the population during this time using both mtDNA and autosomal ancestry data. I stressed the importance of using a combined genetic approach for estimating individual ancestry. The study produced the foundations of a multi-gene ancestry database for Australia, which will continue to grow in the future, allowing a greater understanding of genetic admixture in the Australian population during the World War eras. This database enables objective evaluation of ancestry prediction results for future investigations of historical Australian remains including war dead, missing persons and cold cases.

Significance

The overarching aim of this research was to improve identification outcomes for degraded human remains in Australian forensic casework scenarios where conventional DNA profiling and current forensic MPS techniques may fail to produce meaningful results. Perhaps the biggest emphasis throughout this thesis is on the critical need for a more comprehensive approach to biogeographic ancestry prediction in Australia. In numerous instances throughout this thesis I demonstrate the value of using a combination of mtDNA, Y-chr markers and autosomal SNPs to avoid the risk of misreporting ancestry based on a single biological query. This is particularly important for samples with genetic ancestry originating from two different geographical locations. This concept was further emphasised in the construction of the Historical Australian DNA Database by the inclusion of ancestry information from both mtDNA and autosomal markers. While not performed in this project, the typing of Y-

chromosome SNPs for paternal ancestry can be suggested as another avenue to estimate the genetic ancestry components in the historical Australian population. For the exploration of ancestry testing of highly degraded forensic samples, two novel approaches were developed and evaluated for the potential of increased recovery of forensically relevant intelligence-informative SNPs. Both approaches had been demonstrated throughout the thesis to produce robust results for samples with known biogeographic ancestries, sex and phenotype, and for degraded forensic samples where this information was not known. The successful application of these specialised methods to actual casework samples demonstrates the value of the workflow for future forensic cases (i.e. war dead, cold cases, missing persons) that fail with routine genetic analysis. Currently, it is estimated that approximately 500 sets of unidentified human remains are archived in Australian forensic laboratories (Minister for Justice 2015; Ward 2018). The techniques developed and evaluated in this thesis would be a possible avenue to retrieve intelligence data to narrow the search for a positive identification for these remains, in particular for cases where conventional identification strategies have been unsuccessful.

Currently in Australia, when DNA analysis is requested on degraded/historical remains the first genetic test usually performed is STR typing following DNA quantification (Hartman *et al.* 2015). If STR profiling fails or produces an uninformative result alternative analyses are explored which can consume valuable sample. This may require outsourcing to other laboratories the capability to perform these alternative tasks, which still might have limited success in genotyping degraded samples. Another consideration is whether more specialised techniques such as MPS of STRs and/or SNPs informative for ancestry and phenotype should be used, and what genetic targets should be sought to provide investigative value. This current workflow is costly in time, resources and labour, is logistically difficult and an inefficient use of sample that is already precious in DNA quantity. Furthermore, technical issues still remain for commercial MPS strategies for forensic investigations including expense, inflexibility in marker choice, and the ability to profile highly degraded or compromised remains (Elwick et al. 2018; Gettings et al. 2015). A final concern for the typing of degraded and historical remains in Australia is the lack of a suitable reference population database from which to statistically assess the results from such tests.

The development of the Miniplex in Chapter 2 was designed as an alternative method to DNA quantification for guiding the selection of downstream analysis techniques for degraded samples. Incorporating multiple marker types, across both mtDNA and nuclear DNA with

varying amplicon sizes allows this tool to address two main purposes: to evaluate and triage DNA samples based on degradation of mtDNA versus nuclear DNA, and to provide broad biological profiling that could screen non-probative samples from downstream processing. The prioritisation of samples based on mtDNA and nuclear SNP typing success can streamline workflows, and the comparative evaluation of the genetic targets can guide which profiling strategies should be applied to which samples and thus improve downstream analysis success rates. The suggestion of rapid SNP-based screening tools has been proposed previously (Quintans *et al.* 2004) and has demonstrated utility for streamlining genetic analysis in forensic investigation (Brandstätter *et al.* 2003). However, currently these have only been developed to target a small number mtDNA SNPs, and therefore has limited ability to estimate sample degradation and overall biological profile. The analysis of multiple SNPs across both the mtDNA and nuclear genome with varying amplicon sizes allows for a more detailed evaluation of the length and availability of DNA fragments in the sample. This information can be crucial in deciding whether to apply standard genetic profiling techniques or whether more specialised and sensitive techniques designed for degraded samples with low DNA quantity, such as PCR-based or hybridisation enrichment for MPS, are more likely to retrieve usable genetic profiles for analysis. The Miniplex can therefore help to minimise re-analysis, sample consumption required for multiple tests, and is logistically efficient and practical for the laboratory workflow for compromised forensic samples. Furthermore, as this method makes use of existing laboratory equipment and expertise it can be easily integrated into forensic laboratories.

While the Miniplex has been shown to provide a reliable and accurate assessment of broad biological profile and sample quality from degraded remains it was not designed as a diagnostic tool and is strictly for triaging samples and presumptive purposes only. Hence, the conception, evaluation and application of another, more comprehensive SNP profiling method based on specialised MPS techniques was required and is also described in this thesis. As previously discussed, current PCR-based strategies that apply MPS to forensic samples are limited in their ability to genotype degraded DNA fragments below 150bp, and suffer from substantial locus dropout when encountering PCR inhibitors commonly found in degraded and compromised samples (Gettings *et al.* 2015; Elwick *et al.* 2018). Furthermore, existing MPS panels, whether PCR-based enrichment or hybridisation enrichment, are supplied as pre-made primer/probe sets with pre-designed data analysis and interpretation procedures and do not allow an investigator to tailor genetic analysis on an individual case level. This can result in the abundance of extraneous sequence information that may not hold

any probative value to an investigation. In conjunction, the 'big data' generated puts pressure on the demand for computational resources, cost and specialised expertise for best practices for data storage and management. Ethical concerns for the application of DNA phenotyping (including ancestry, hair and eye colour prediction) using large SNP panels with upwards of 200 genetic markers has also sparked discussion amongst forensic and legal practitioners (Scudder *et al.* 2018). As more markers in the human genome are sequenced, more attributes of the donor could be revealed that are not important for the investigation and thus safeguarding genetic privacy of the donors and their families is a challenge for emerging forensic intelligence testing. This poses the question of whether we should collect data for hundreds of genetic markers because new technologies allow us to do so, or whether we should be limiting analysis to the minimum number of markers relevant to an investigation to maintain 'genetic privacy'. Thus, the successful development, and application of a novel 'modular' hybridisation enrichment approach (that can be easily customised by use of individual baits to answer specific questions for degraded remains) is a new tool through which more tailored genetic analysis using MPS can be potentially utilised. While based on MPS technologies, the customisable enrichment strategy developed and applied here does not seek to exploit the capability to sequence an excessive number of genetic markers for forensic intelligence. The combination of multiple marker types in the panel is able to resolve ancestry components more accurately than using a single marker type alone. The ability of the enrichment panel to retrieve genetic information from degraded teeth and fragmentary bone with a PMI of over 70 years shows the benefit of this technique for typing and making inferences of ancestry, paternal lineage, sex and phenotype for degraded casework DNA samples where conventional DNA profiling may fail. However, it should be noted that regardless of the approach for ancestry prediction, the implications on the social and cultural identity of individuals, especially those with Indigenous heritage, cannot be ignored and has been the focus of recent commentary surrounding genetic ancestry testing (Kowal & Jenkins 2016; Booth 2018; Watt & Kowal 2018). In particular, is the concern for ancestry testing to determine 'Aboriginality', especially for Australian Aboriginals where events such as European admixture and the Stolen Generations has diluted and complicated Australian Aboriginal ancestry. As a result, Aboriginality is determined by documented ancestors with pre-confirmed acceptance by an Aboriginal community, and cultural affiliation to a tribe. Yet more Australians than ever are identifying themselves as Indigenous Australians, and with it the question of who really 'is and isn't' Aboriginal Australian has been asked (Markham & Biddle 2018; Watt & Kowal 2018). The idea of using a genetic ancestry test to confirm Aboriginality has been both raised and debated on the grounds of the lack of Australian

Aboriginal reference autosomal SNP data for accurate comparisons, and the view from Indigenous communities that an individual's DNA does not reflect cultural identity within a community (Kowal & Jenkins 2016; Booth 2018). In fact, Rachael Hocking - a woman with known Australian Aboriginal descent and strong cultural affiliation to her Walpiri nation - revealed no Indigenous ancestry when taking an ancestry test from DNA Tribes, one of only two providers who offer an 'Aboriginality test' for Australians (Booth 2018). She highlighted the plight of Aboriginal Australians who suffered from the events of the Stolen Generation and are 'looking for closure' through this genetic ancestry test, who are told they do not identify as Aboriginal Australian (Booth 2018). Instead, Aboriginal elders state that 'cultural knowledge and experience of living Black', community acceptance and connection to country are the most important factors in determining Aboriginality (Kowal & Jenkins 2016; Watt & Kowal 2018).

As the forensic intelligence testing presented in this thesis has demonstrated the potential to genotype markers from historical human samples in Australia, a final element to the workflow was addressed through the compilation of the Historical Australian DNA Database (HADD). This database facilitates objective assessment of ancestry testing results from historical remains. Ongoing efforts to retrieve and repatriate historical war dead remains from World War I (WWI) and II (WWII) battlefields has resulted in a large number of unidentified human remains. These require assignment to 'country of origin' in the first instance before more comprehensive individual identification strategies can occur. In many cases, the compromised, skeletal and fragmentary nature of the remains, which are now >70 years old, means DNA analysis techniques such as ancestry, maternal and paternal lineage and phenotype prediction can be the only source of this information. However, the absence of a suitable population database to use as a reference population places some doubt on how to analyse, interpret and report ancestry results. The genetic ancestries that actually contributed to the Australian population at this time is therefore uncertain. Two features of the HADD make it an extremely useful resource for interpreting ancestry results from historical Australian remains. Firstly, the targeted inclusion of individuals who represent the Australian population before the major waves of migration post WWII allows for a more accurate estimation of the genetic components of the population during the World War eras. Secondly, the database combines multiple genetic marker types (mtDNA and autosomal SNPs), reducing the risk of misclassifying an individual's genetic ancestry - particularly those with admixed ancestry. The detection of ancestry components derived from mtDNA and autosomal DNA originating from different continental regions in two of the samples was a

demonstration of this concept. It is therefore suggested that any ancestry analysis of historical Australian human remains should include multiple marker types for more accurate ancestry assignment. The need for representative population databases to generate statistical support for mtDNA and intelligence testing for Australian human remains has previously been voiced, yet such progress in Australia is so far quite limited (Ward 2016). As the HADD continues to expand with future efforts, it will provide a valuable objective genetic resource to add statistical support for ancestry predictions based on mtDNA and autosomal DNA obtained from recovered war dead.

Broader Applications

While this thesis primarily focused on the application of forensic intelligence testing to degraded historical remains, it is also recognised that this research has many possible broader applications to other situations that require genetic investigation.

Modern Crime Scene Samples

Investigations of present-day criminal cases can be impeded if there is no match between a DNA profile obtained from probative crime scene evidence to existing profiles in a criminal database. The absence of any eye witnesses can also preclude identification of a suspect. Even when available, inaccuracies regarding eye witness accounts can make this evidence unreliable (Spinney 2008; Kayser & Schneider 2009; Wise *et al.* 2014).The development of tests to infer ancestry and pigmentation in skin, hair and eyes is a new era of forensic intelligence gathering where DNA is now not only used by enforcement agencies solely for matching a crime scene sample with a suspect or database profile. If there are no profile-to-profile matches, and no witnesses, intelligence information may provide guidance in the targeted DNA sampling of possible persons of interest. This intelligence information could also help to corroborate eyewitness statements where available (Jacobson 2005; Johnston 2006). Narrowing suspect pools by identifying characteristics most likely exhibited by the perpetrator allows investigators to expend resources on the highest probability leads. This was demonstrated most notably in the Madrid bombing incident where ancestry testing on crime scene samples determined the most likely ancestry of donor/suspect to be North African. These results eventually lead to the DNA profiling of a relative's DNA sample that resulted in a match with a known Algerian terrorist who was consequently arrested (Phillips *et al.* 2009).

Given the ability of the Miniplex and custom hybridisation enrichment panel to type samples with limited or degraded DNA, the methods described in this thesis have merit in analysing problematic crime scene samples that may fail with standard PCR-based STR typing such as telogen hairs (Edson *et al.* 2013), trace samples (Hanssen *et al.* 2017) and touch DNA (Martin *et al.* 2018). The benefit is not only in the technology to retrieve SNPs from such samples, but in the typing of many forensically relevant SNP classes that can determine overall biological profile regarding paternal lineage and sex, biogeographic ancestry, and phenotype in a single assay. The Miniplex can aid in the careful selection of samples that will likely yield the most genetic information. Additionally, the Miniplex can provide a broad mtDNA haplogroup and indicate whether mtDNA control region analysis or whole mtDNA genome sequencing may offer genetic information to assist in identification. The broad inferences offered by the Miniplex may help screen samples and eliminate those which are not probative to the investigation. SNP typing techniques described and evaluated in this thesis may also have value in the analysis of historic crime scene samples from cold cases to obtain forensic intelligence data where not previously possible, especially for DNA samples which might have degraded in storage over time. Fragmented and partial human remains where dental, fingerprint and anthropological analysis is inconclusive or cannot proceed due to a lack of diagnostic features such as teeth and cranial fragments (and where STR typing produces no matches) may also benefit from such techniques. Intelligence data from these remains may provide clues to the characteristics of the person in the absence of other biological information or could complement other biological findings (such as anthropological sex).

Archaeology/Ancient Human DNA Studies

Studying the DNA of ancient individuals through time and across space provides insight into the history and lifestyles of past human populations (Haak *et al.* 2015; Lipson *et al.* 2017; Nielsen *et al.* 2017). Samples commonly used for these analyses include bones, teeth and mummified tissue that are recovered from archaeological sites and ancient burials. Analysis of ancient DNA is difficult due to poor preservation, where surviving DNA molecules are both degraded and damaged, and can exist in vanishingly small amounts (Hofreiter *et al.* 2015). As DNA fragments become shorter over time by natural degradative processes, the amount of original endogenous DNA that can be recovered and analysed is reduced, in some cases as low as 1% of the original DNA content (Carpenter *et al.* 2013). Major technological developments in the field of MPS has allowed unprecedented understanding of the genomic

variation and genetic characterisation of past human populations that wasn't previously possible with traditional PCR processes (Lipson *et al.* 2017; Nielsen *et al.* 2017). Therefore, the customisable hybridisation enrichment and data analysis strategy developed and evaluated throughout this thesis can also be used to retrieve and analyse genetic information regarding ancestry, paternal lineage and hair and eye colour from such samples. However, the Miniplex may also be used in the first instance to select appropriate samples for taking through the MPS workflow owing to the ability to comparatively assess varying amplicon sizes across mtDNA and nuclear DNA. The power to provide inferences of broad biological profiling using the Miniplex may also be useful to screen samples based on the purposes of the ancient DNA study e.g., aiming to analyse individuals only from a particular mtDNA or Y-chr lineage (Malyarchuk *et al.* 2010). Therefore, the Miniplex can also help to reduce the labour and costs of analysing many samples, some of which may not be relevant to the aims of the study.

Contemporary Australian Population Data

As discussed throughout this thesis, mtDNA analysis in Australia can be complicated by the absence of any suitable reference population database that estimates the frequency of haplotypes and maternal ancestry groups in the Australian population. While the EMPOP mtDNA database has made efforts towards the compilation of global forensic-standard mtDNA population data, it is still over-represented by Western Eurasian populations with smaller datasets of East Asian, South East Asian, and Sub-Saharan African population groups (Parson & Dur 2007). The use of such a database for Australian cases may skew results and can confound the interpretation of mtDNA testing in Australia.

The HADD was constructed specifically to represent the Australian population pre-1945 for direct application to cases involving long-term missing persons during this time and ancestry testing of historical human remains including war dead. However, a suitable reference population database for a modern Australian population also does not exist. It is currently estimated that there are 500 unidentified human remains and 2000 long-term missing persons in Australia (Ward 2018). The Victorian Institute of Forensic Medicine has demonstrated success at using a state-wide DNA identification program for missing persons using mtDNA and mtDNA databases, however no other approach has been implemented in other jurisdictions (Hartman et al. 2015). Proposal for a national DNA-led program that incorporates STRs as well as mtDNA, Y-chromosome analysis and SNPs for intelligence testing has been advocated recently to assist in the identification of missing persons across

Australia (Ward, 2018; 2016). However, with these suggestions comes the premise that reference population databases using these markers in a contemporary Australian population are needed to provide statistical support for results (Ward 2016). The HADD could therefore contribute to a reference population database that estimates the frequency of mtDNA haplotypes, haplogroups and ancestry-informative markers in a modern Australian population given that donors are not only living individuals representative of the historical population but also currently resident in Australia. Since no modern Australian reference population exists for this purpose, it could serve as the basis for such a database and can be built upon with more samples representative of the contemporary population for present-day missing persons cases. There is also the possibility of using the database alongside the EMPOP database to provide statistical support for results from both a national and international population assessment.

Limitations and Recommendations for Future Directions

The investigations performed throughout this thesis were limited by a number of factors that are either specific to the type of data I generated, linked to the conceptual approach used or are generally related to the field of forensic genetics and degraded/damaged DNA. In the proceeding sections I discuss these issues and propose new opportunities and future directions that could help overcome them.

Database Sample Size and Quality Control

A large sample size through the collection of ~800 public donor samples was collected as part of the HADD project but time constraints during this thesis limited the analysis to only 259 samples for the mtDNA control region and autosomal ancestry SNPs. Large mtDNA datasets are needed to evaluate haplotype and haplogroup frequencies within a population due to the strict maternal inheritance, in contrast to autosomal markers that are inherited with recombination from both parents (Budowle *et al.* 2003). With ongoing efforts, more samples will be analysed for expansion of the database to increase the strength and utility of the dataset. However, at present the database should be considered for forensic casework with caution due to issues relating to small sample size. Despite addressing the uncertainty of frequency values (as a result of sample size) through the calculation of confidence intervals (CI), it is acknowledged smaller sample sizes generate wider confidence intervals (Gauthier & Hawley 2015). This leads to a large window in which the true frequency of ancestry

groups in the Australian population resides and thus any statistical interpretation of mtDNA or autosomal ancestry from unknown remains generated should account for this uncertainty. Eventually, genotyping additional samples for the database will lead to smaller CI's and theoretically increase the reliability of the frequency estimates calculated, thus increasing the power and value of the database for forensic purposes.

Although no mention was made in regard to ethnicity or specific ancestries targeted in the information material supplied to the public, individuals with non-European ancestry who represent the Australian population pre-1945 might not culturally identify as Australian and may be deterred from donating a DNA sample. This may reduce the chance of detecting any non-European ancestry in the dataset and bias frequency estimates and confidence intervals despite those individuals truly reflecting the Australian population during this time. While every effort was made to guard against this potential bias during sampling, it was outside my scope to control.

Although internal quality control (QC) measures were implemented to ensure only high-quality haplotypes were being analysed and interpreted for entry into the historical database, no independent external QC was incorporated into the research. The value of external QC has been recognised in previous mtDNA studies and in the compilation of the EMPOP database to ensure uniformity of mtDNA sequencing and consistency in the nomenclature of results across forensic laboratories (Parson *et al.* 2004; Parson & Dur 2007). Current standards for mtDNA data QC include the use of software tools within EMPOP:

- a phylogenetic approach for detecting potential errors in the sequencing, interpretation and transcription in a dataset using a quasi-median network (QMN) analysis (available in the EMPOP *NETWORK* application).
- Haplogroup affiliation using a maximum likelihood approach *EMMA*, inbuilt software within the EMPOP query function that considers private mutations and phylogenetic information from PhyloTree to assign haplogroups

While haplogroup affiliation using EMMA was implemented in this thesis by analysing haplotypes in EMPOP, it is also recommended by Parson et al. (2007a) that the developers of EMPOP provide an external QC check on submitted samples to ensure high quality data is included in EMPOP if desired (Parson & Bandelt 2007). QMN analysis of datasets can detect unusual or previously unobserved length or sequence variants based on previously known

phylogenetic information and prompt the investigator to review raw sequencing data. This QC check has been implemented for both internal and external QC in previous studies (Chaitanya *et al.* 2016; Turchi *et al.* 2016), where investigators have submitted their datasets to EMPOP for external review of the haplotype data. The use of EMMA also aims to standardise mtDNA nomenclature across studies of mtDNA variation and is also recommended to be performed externally as another measure for data quality control. Successful external QC is documented by an accession number that acts as a unique identifier for the dataset and confirms completion of the data review. The mtDNA data in this thesis has not been subjected to external review since submission to EMPOP was not an initial aim of the studies.

Despite these limitations, this thesis has presented the first step towards the construction of a large reference population database for ancestry determination of historical war dead and missing persons.

Sample/DNA preservation and the Recovery of Genetic Information

For some degraded and casework samples tested in my research, I was unable to retrieve sufficient genetic data to make inferences of intelligence information. In many instances DNA extracts processed previously in our laboratory (up to ten years in some cases) were used for analysis due to the complexity of obtaining degraded and casework samples. This can explain the results in Chapter 2, where SNP typing using the Miniplex was not shown to be significantly more sensitive – despite much shorter amplicon lengths - than previous STR typing which had occurred years earlier when extracts where freshly produced in Higgins *et al.* (2015). Mock degraded samples (usually pristine control DNA samples that are mechanically fragmented) were not desired in this study since they don't truly reflect the natural environmental degradation and damage processes that occur within biological material post-mortem (Budowle *et al.* 2009). However, it is acknowledged that mock 'degraded' samples with known genotypes would have been useful for a further exploration of the Miniplex and the hybridisation enrichment panel for the capacity to reliably type and infer the ancestry of samples which could represent degraded DNA. The level of endogenous DNA available for analysis in highly degraded samples such as those analysed in this thesis may be extremely low. The recovery and genotyping of this DNA is further dependent on the extraction process, and the length and conditions of storage of extract thereafter. Environmentally challenged, degraded or damaged samples containing low DNA

concentrations are more susceptible to genotyping failure when stored long-term in polypropylene tubes due to repeated freeze-thawing (Davis *et al.* 2000), adherence of DNA to the tubes (Gaillard & Strauss 1998), evaporation or denaturation (Gaillard & Strauss 2001). These factors may have influenced the success of the methods to retrieve and genotype the DNA from these samples. Re-extraction of the degraded and casework samples would be ideal to assess the capability of the genotyping methods on fresh extracts and may improve on the success of SNP recovery demonstrated within this thesis. Preferably, multiple extractions per sample would be performed, pooled and concentrated to a small volume for analysis for improved recovery and depth of coverage across the SNPs.

Additionally, a higher sequencing effort may lead to increased coverage over the SNP targets or detect more loci. These techniques may have the potential to recover more targets to a higher read depth for improved inferences of ancestry, phenotype, mtDNA and Y-chr haplogroups. Performing whole mtDNA genome enrichments already established in our laboratory would also have been useful as another means to assess sample degradation for DNA availability and to compare to nuclear enrichment data (Templeton *et al.* 2013). A further comparison of the Miniplex and hybridisation enrichment techniques to standard CE technologies, as well as current commercial PCR multiplexes for MPS of forensic samples would have also been useful to assess the performance of the techniques in this thesis against what is currently available.

Sample Contamination/DNA Mixtures

This study did not explore in great detail the potential for contamination of samples with exogenous human DNA. The use of bi-allelic SNPs to determine and resolve the presence of a mixture is challenging (Gill 2001). For tri- and tetra-allelic SNPs, mixtures can be detected when more than two alleles are observed at a locus (Phillips *et al.* 2015). For haploid markers, mixtures can be detected when two different alleles are observed in the analysis (Bose *et al.* 2018). While no mixed profiles were detected by the use of the multi-allelic SNPs or haploid markers in any case throughout this thesis, the capability of the methods to detect mixtures was not determined beforehand. It would be valuable to analyse mixtures of known genotypes in pre-determined ratios to assess the ability to detect DNA contamination and resolve mixtures. This is especially important for female:female mixtures, given that only a few haploid mtDNA markers will be obtained using the Miniplex, and the tri- or tetra-allelic autosomal markers are the only source for mixture detection in the enrichment method.

For these purposes, extra multi-allelic markers could be easily included in further revisions of the enrichment panel. To accurately evaluate how the methods cope with and detect DNA mixtures is useful knowledge to have considering forensic samples can be susceptible to DNA contamination. However, extending this analysis to degraded DNA mixtures may be experimentally challenging.

Classification Approaches

The accuracy of ancestry assignment relies on a number of factors, including the use of appropriate classification methods. Three different classifiers were used for estimating ancestry components throughout this thesis. STRUCTURE is currently the classification method of choice but it is computationally demanding and complex for large sample sizes and SNP sets (Cheung *et al.* 2017). Principle Component Analysis (PCA) only provides a graphical representation of the genetic variation using bi-allelic loci, and the use of the online Bayesian classifier Snipper (Santos, C. *et al.* 2016b) is currently limited in accurately accounting for admixture (Cheung *et al.* 2017), so these approaches also come with their own limitations. Not all these classification systems are suitable for analysis alone (especially when dealing with admixed individuals), hence a combination of the three methods was used throughout this thesis to represent the data and assess ancestry assignments. Perhaps the most significant hurdle of these classification algorithms in assigning ancestry is the assessment of samples with ancestry admixture (Cheung *et al.* 2018). It is also not currently clear how well the panel and classification workflows in this thesis handle admixed samples, since only a small number of samples showing admixture signals were analysed. Comparisons of standard classifiers, as well as alternative algorithms have recently been explored on a dataset of 142 biogeographic ancestry SNPs (Cheung *et al.* 2017, 2018). Cheung et al (2018) showed that STRUCTURE and a genetic distance algorithm (GDA) were most successful in detecting and resolving ancestry components in admixed samples. Neither method outperformed the other across all simulated admixture complexity and ratios. However, since ancestry assignment is also dependant on the SNP panel, these results cannot be extrapolated to the marker sets used in this thesis. In this case, an exploration of alternative classifiers such as GDA would be beneficial to perform with these panels for further evaluation and refinement of the ancestry analysis workflow. Analysis of samples considered unadmixed, as well as populations with high ancestry admixture signals should be included for an overall assessment of the ancestry assignment methods developed and applied throughout this research. Although admixed individuals from populations studies in 1000 Genomes and the HGDP datasets are a good

starting point, admixed individuals representative of a modern multicultural Australian population would be ideal to analyse for feasibility of the method for real-life scenarios. Alternatively, simulations to create artificially admixed samples such as those performed in Cheung et al (2018) (however with the addition of Oceanian ancestry considering Australia's Aboriginal population history) could be performed to assess the prediction accuracy and capability of the workflow to detect and resolve admixture in an Australian population.

It should also be noted that the use of SNP panels and classification methods for the detection of admixture can be further confounded by sample degradation. DNA degradation and damage could result in a loss of the SNP data informative for ancestry admixture, further complicating the ability to assess admixture in degraded samples. Existing studies that assessed the capability of SNP panels and classification systems for admixture detection, utilised admixed sample data where complete profiles are obtained and analysed (Halder *et al.* 2008; Nievergelt *et al.* 2013; Phillips *et al.* 2014; Cheung *et al.* 2018). This does not account for the event of both ancestry admixture and partial profiles from degraded DNA samples. A possible avenue to address this is to generate genotypes from known admixed samples and randomly simulate missing data to varying degrees to assess how partial profiles influence the assignment of admixed samples, and how different classifiers handle this data using the SNP panels in this thesis.

However, regardless of the classification method used, the most critical factor in accurately determining biogeographic ancestry of unknown samples is the size, coverage and quality of reference population datasets available for comparison. Predicting ancestry from regions underrepresented due to irregular global coverage in publicly available datasets remains a concern for geographical locations such as Oceania. Currently available autosomal SNP data that is routinely used as a Oceanian reference population set in forensic ancestry SNP analysis only includes a small number of samples from Papua New Guinea and Bouganville (17 Papuan from New Guinea and 11 Melanesian from Bouganville) (Cann *et al.* 2002). In the context of population databases, this could be considered an inadequate sample size for estimating allele frequencies, particularly for more rare alleles, selecting informative SNP loci, and is not representative of population variability across the whole Oceanian region which also encompasses Australian Aboriginal, Micronesian and Polynesian populations. Despite this, a study using an autosomal ancestry SNP panel (Pacifiplex) for the inference of Oceanian ancestry showed that Fijian and Australian Aboriginal samples from Northern Territory and Western Australia clustered with the Oceanian reference dataset, and thus was considered an

appropriate test for Australian Aboriginal samples. However, Micronesian and Polynesian test samples (populations not included in the Oceanian reference population dataset) had a majority or complete cluster membership to the East-Asian reference dataset (Santos, Carla *et al.* 2016a). Future studies including greater sample numbers as well as improving the geographical coverage of Oceanian sub-populations, and other under-represented geographical regions, will result in a greater understanding of the human population history of these areas, and in a more suitable dataset that can be routinely used for interpreting results from forensic ancestry SNP typing techniques.

Evaluation of Alternative Marker Types

Standard mtDNA, autosomal and Y-chr SNPs already widely accepted and well established in the forensic community were utilised in this study for ancestry and phenotype analysis. However, more recent studies are showing that other emerging markers and more recent panels may also be a valuable source of biological information or improve analysis in the future (Kidd, KK *et al.* 2014; Phillips *et al.* 2015; Chaitanya *et al.* 2018). Given the possibility for customisation of the hybridisation enrichment baits in this thesis, these alternative marker types can be included in further revisions of the panel or could be used alone if desired.

The use of microhaplotypes (multiple SNPs closely linked) has been advocated for and shown to allow kinship and lineage testing, mixture detection and ancestry inference (Kidd, JR *et al.* 2011; Pakstis *et al.* 2012; Kidd, KK *et al.* 2013; Kidd, KK *et al.* 2014; Bose *et al.* 2018). These sequence stretches of no more than 200 bp contain two or more linked SNPs do not recombine, thereby acting as multi-allelic haplotype markers (Kidd, KK *et al.* 2013). The multiple alleles of these haplotypes can be more informative than bi-allelic SNP loci for determining lineages, individual identification and for ancestry inference with the added power of detecting DNA mixtures. Genotyping these microhaplotypes is becoming more accessible with advancements in MPS technologies. As a result, these markers are being increasingly explored for their application to forensic investigations (Kidd, KK *et al.* 2014). While microhaplotypes hold potential for improving inferences of ancestry and for detecting mixtures, validation studies are limited. Further, the lack of implementation across forensic laboratories means forensic population data that can be used for statistically evaluating and reporting results is virtually absent. Future improvements in the selection, genotyping and interpretation of microhaplotypes with associated validation studies and population databases

may make microhaplotype markers appealing for inclusion into the custom hybridisation enrichment panel in any further revisions. Typing microhaplotypes via standard CE technologies is laborious and time consuming since each SNP within the microhaplotype requires genotyping via individual primer sets (Kidd, KK *et al.* 2014), and is therefore not desirable for implementing into SNaPshot SNP panels

The genotyping of multi-allelic SNP loci (tri- and tetra-allelic SNPs) has been identified as a means to overcome the limitations of bi-allelic SNPs in detecting and resolving DNA mixtures, and for increasing discrimination power of identification SNPs (Phillips *et al.* 2015; Bose *et al.* 2018). A small number of tetra-allelic SNPs have also been identified for their potential use in ancestry inference, albeit with less well-distributed variation among population groups than standard ancestry informative bi-allelic SNPs (Phillips *et al.* 2015). Multi-allelic SNPs should not be considered as replacement marker sets to existing forensic SNP panels, however they can form a supplementary marker type for inclusion into analysis workflows for evaluating mixtures and ancestry inferences (Phillips *et al.* 2015). Since evaluating mixture detection was not a main aim of the research in this thesis, very few multi-allelic SNPs were included in the hybridisation enrichment panel design. Further evaluations and revisions of the hybridisation enrichment strategy presented in this thesis may also include more tri- and tetra-allelic markers for the purposes of mixture detection and resolution, and for the potential improvement of ancestry inference.

The HIrisPlex phenotype SNPs validated for forensic use were included and analysed in the SNP panels in this thesis for predicting hair and eye colour (Walsh *et al.* 2014). Since the development of the hybridisation enrichment panel, the 'HIrisPlex-S' panel now exists which types a further 17 SNPs for a total of 41 markers across the HIrisPlex and HIrisPlex-S panels for predicting hair, eye and skin pigmentation (Chaitanya *et al.* 2018). The HIrisPlex-S panel is capable of predicting five skin colour categories from 'very pale' (76% prediction accuracy) to 'dark to black' (99% prediction accuracy) and is currently the only forensically validated tool for predicting skin colour. The prediction model for skin colour using the HIrisPlex-S system includes reference database samples from global populations from the HGDP-CEPH dataset and a US-based study including individuals born outside of the US (Walsh *et al.* 2017). This addresses previous concerns with the HIrisPlex model where only European samples were used (Bulbul & Filoglu 2018). However, hair and eye colour prediction models are still built on reference databases dominated by European samples with only a relatively small number of samples from other global populations included

(https://hirisplex.erasmusmc.nl/pdf/hirisplex.erasmusmc.nl.pdf). Nonetheless, the combined HIrisPlex and HIrisPlex-S panel is now capable of predicting all three human pigmentation traits which can collectively provide valuable visual descriptors of an individual to aid in forensic investigations and could be useful markers to include in the hybridisation enrichment panel.

Reporting of Forensic Data from MPS Approaches

The application of MPS for forensic analysis has prompted many studies that have explored the recovery of genetic information from hundreds of markers (Gettings *et al.* 2015; Churchill *et al.* 2016; Eduardoff *et al.* 2016; Apaga *et al.* 2017; de la Puente *et al.* 2017; Bose *et al.* 2018; Ma *et al.* 2018). However, the field is still in its infancy with many concerns and considerations left to be addressed, particularly for its application to degraded samples. Studies using MPS have been mostly performed on control DNA samples, or mock degraded samples that are fragmented and size selected from pristine control DNA but may not necessarily reflect the natural degradation processes (where the DNA is present in both low quantity and quality). Furthermore, published PCR-based and enrichment technologies are either incapable of distinguishing PCR duplicates, or do not remove PCR duplicates prior to genotype calling (Bose *et al.* 2018; Shih *et al.* 2018). Duplicate removal was performed in this thesis as a means to minimise errors during genotype calling with the acknowledgement that read depths over targets will be substantially reduced in samples with high clonality. Read depth thresholds applied in previous studies (with no duplicate removal) range from 2x (Elwick *et al.* 2018), 10x (Bose *et al.* 2018) and 50x (Gettings *et al.* 2015), but a formal consensus of read depth thresholds for forensic purposes has not been reached. Consequently, in this thesis no formal read depth threshold was adopted for reporting genotypes. Instead, analysis was performed at three different read depth thresholds to observe the effect on resulting interpretations and predictions. It was found that increasing the minimum read depth threshold greatly reduced the amount of available data in degraded samples and resulted in predictions with weak support. Detailed studies evaluating MPS approaches on true degraded samples and further validation studies using replicates to assess the impact of read depth thresholds on genotyping accuracy are needed to empirically determine appropriate read depth thresholds and establish best practices for analysing, interpreting and reporting MPS data for forensic casework. One such example can be found in the experimental design in Higgins *et a.l* (2015), where freshly extracted human teeth with known STR profiles were buried for varying periods of time up to 16 months and STR genotyped to examine the effects

of degradation of various tooth tissues on the retrieval of DNA (Higgins *et al.* 2015). A similar experiment involving MPS would be an avenue to explore the effects of authentic DNA degradation on MPS sequencing technologies and whether they are capable of generating reproducible and accurate genotyping results and inferences of ancestry at varying read depth thresholds. Of particular importance, would be assessing this effect at low read depth thresholds such as 2x or 5x for application to degraded DNA with very low amounts of DNA. Additionally, analytical read depth thresholds will differ depend on what method of target enrichment is used (e.g PCR amplicon sequencing vs hybridisation enrichment), given that amplicon sequencing involves the generation of PCR clones which result in read duplicates that cannot be distinguished via bioinformatic processing. On the other hand, hybridisation enrichment allows the filtering of read duplicates which can inflate read depths over target loci, such that only unique sequencing reads are used in genotype calling. For this reason, comparative studies between amplicon sequencing and hybridisation enrichment on mock degraded and authentic degraded samples will be needed to understand the different requirements for setting analytical thresholds for the interpretation of MPS data.

Concluding Remarks

The overarching aim of this research was to explore and develop new tools to increase the likelihood of drawing inferences regarding ancestry and phenotype from degraded human DNA for forensic investigations in Australia. This has been achieved through the development, evaluation and application of novel genotyping systems, and in the collection of population data from an historical Australian populace for use in ancestry determination. Throughout this research I encountered challenges as well as additional areas worthy of exploration that could not be addressed during my candidature yet would serve as interesting points of investigation for future studies. The knowledge and techniques gained in this research have been applied to degraded DNA and casework extracts, and has shown value in enabling the inference of forensic ancestry and phenotype data from degraded human remains. Overall, this research has presented new considerations and avenues from which to gain forensic intelligence information that could be vital in directing investigations of identity.

References

Apaga, D.L.T., Dennis, S.E., Salvador, J.M., Calacal, G.C. & De Ungria, M.C.A. 2017. Comparison of Two Massively Parallel Sequencing Platforms using 83 Single Nucleotide Polymorphisms for Human Identification, *Sci Rep*, 7, 398.

Booth, A. 2018, 'Indigeneity and DNA ancestry tests', *The Saturday Paper*, October 20, 2018,

Bose, N., Carlberg, K., Sensabaugh, G., Erlich, H. & Calloway, C. 2018. Target capture enrichment of nuclear SNP markers for massively parallel sequencing of degraded and mixed samples, *Forensic Sci Int Genet*, 34, 186-96.

Brandstätter, A., Parsons, T.J. & Parson, W. 2003. Rapid screening of mtDNA coding region SNPs for the identification of west European Caucasian haplogroups, *Int J Legal Med*, 117, 291-8.

Budowle, B., Allard, M.W., Wilson, M.R. & Chakraborty, R. 2003. Forensics and mitochondrial DNA: applications, debates, and foundations, *Annu Rev Genomics Hum Genet*, 4, 119-41.

Budowle, B., Eisenberg, A.J. & van Daal, A. 2009. Validity of Low Copy Number Typing and Applications to Forensic Science, *Croat Med J*, 50, 207-17.

Bulbul, O. & Filoglu, G. 2018. Development of a SNP panel for predicting biogeographical ancestry and phenotype using massively parallel sequencing, *Electrophoresis*, 39, 2743-51.

Cann, H.M., de Toma, C., Cazes, L., Legrand, M.F., Morel, V., Piouffre, L., Bodmer, J., Bodmer, W.F., Bonne-Tamir, B., Cambon-Thomsen, A., et al. 2002. A human genome diversity cell line panel, *Science*, 296, 261-2.

Carpenter, Meredith L., Buenrostro, Jason D., Valdiosera, C., Schroeder, H., Allentoft, Morten E., Sikora, M., Rasmussen, M., Gravel, S., Guillén, S., Nekhrizov, G., et al. 2013. Pulling out the 1%: Whole-Genome Capture for the Targeted Enrichment of Ancient DNA Sequencing Libraries, *Am J Hum Genet*, 93, 852-64.

Chaitanya, L., Breslin, K., Zuñiga, S., Wirken, L., Pośpiech, E., Kukla-Bartoszek, M., Sijen, T., Knijff, P.d., Liu, F., Branicki, W., et al. 2018. The HIrisPlex-S system for eye, hair and skin colour prediction from DNA: Introduction and forensic developmental validation, *Forensic Sci Int Genet*, 35, 123-35.

Chaitanya, L., van Oven, M., Brauer, S., Zimmermann, B., Huber, G., Xavier, C., Parson, W., de Knijff, P. & Kayser, M. 2016. High-quality mtDNA control region sequences from 680 individuals sampled across the Netherlands to establish a national forensic mtDNA reference database, *Forensic Sci Int Genet*, 21, 158-67.

Cheung, E.Y.Y., Gahan, M.E. & McNevin, D. 2017. Prediction of biogeographical ancestry from genotype: a comparison of classifiers, *Int J Legal Med*, 131, 901-12.

Cheung, E.Y.Y., Gahan, M.E. & McNevin, D. 2018. Prediction of biogeographical ancestry in admixed individuals, *Forensic Sci Int Genet*, 36, 104-11.

Churchill, J.D., Schmedes, S.E., King, J.L. & Budowle, B. 2016. Evaluation of the Illumina® Beta Version ForenSeq™ DNA Signature Prep Kit for use in genetic profiling, *Forensic Sci Int Genet*, 20, 20-9.

Davis, D.L., O'Brien, E.P. & Bentzley, C.M. 2000. Analysis of the degradation of oligonucleotide strands during the freezing/thawing processes using MALDI-MS, *Anal Chem*, 72, 5092-6.

de la Puente, M., Phillips, C., Santos, C., Fondevila, M., Carracedo, Á. & Lareu, M.V. 2017. Evaluation of the Qiagen 140-SNP forensic identification multiplex for massively parallel sequencing, *Forensic Sci Int Genet*, 28, 35-43.

Edson, J., Brooks, E.M., McLaren, C., Robertson, J., McNevin, D., Cooper, A. & Austin, J.J. 2013. A quantitative assessment of a reliable screening technique for the STR analysis of telogen hair roots, *Forensic Sci Int Genet*, 7, 180-8.

Eduardoff, M., Gross, T.E., Santos, C., de la Puente, M., Ballard, D., Strobl, C., Børsting, C., Morling, N., Fusco, L., Hussing, C., et al. 2016. Inter-laboratory evaluation of the EUROFORGEN Global ancestry-informative SNP panel by massively parallel sequencing using the Ion PGM™, *Forensic Sci Int Genet*, 23, 178-89.

Elwick, K., Zeng, X., King, J., Budowle, B. & Hughes-Stamm, S. 2018. Comparative tolerance of two massively parallel sequencing systems to common PCR inhibitors, *Int J Legal Med*, 132, 983-95.

Gaillard, C. & Strauss, F. 1998. Avoiding adsorption of DNA to polypropylene tubes and denaturation of short DNA fragments, *Technical Tips Online*, 3, 63-5.

Gaillard, C. & Strauss, F. 2001. Eliminating DNA loss and denaturation during storage in plastic microtubes, *Am Clin Lab*, 20, 52-4.

Gauthier, T.D. & Hawley, M.E. 2015, 'Chapter 5 - Statistical Methods', in BL Murphy & RD Morrison (eds), *Introduction to Environmental Forensics (Third Edition)*, Academic Press, San Diego, pp. 99-148.

Gettings, K.B., Kiesler, K.M. & Vallone, P.M. 2015. Performance of a next generation sequencing SNP assay on degraded DNA, *Forensic Sci Int Genet*, 19, 1-9.

Gill, P. 2001. An assessment of the utility of single nucleotide polymorphisms (SNPs) for forensic purposes, *Int J Legal Med*, 114, 204-10.

Haak, W., Lazaridis, I., Patterson, N., Rohland, N., Mallick, S., Llamas, B., Brandt, G., Nordenfelt, S., Harney, E., Stewardson, K., et al. 2015. Massive migration from the steppe was a source for Indo-European languages in Europe, *Nature*, 522, 207.

Halder, I., Shriver, M., Thomas, M., Fernandez, J.R. & Frudakis, T. 2008. A panel of ancestry informative markers for estimating individual biogeographical ancestry and admixture from four continents: utility and applications, *Human Mutat*, 29, 648-58.

Hanssen, E.N., Lyle, R., Egeland, T. & Gill, P. 2017. Degradation in forensic trace DNA samples explored by massively parallel sequencing, *Forensic Sci Int Genet*, 27, 160-6.

Hartman, D., Benton, L., Spiden, M. & Stock, A. 2015. The Victorian missing persons DNA database – two interesting case studies, *Aust J Forensic Sci*, 47, 161-72.

Higgins, D., Rohrlach, A.B., Kaidonis, J., Townsend, G. & Austin, J.J. 2015. Differential Nuclear and Mitochondrial DNA Preservation in Post-Mortem Teeth with Implications for Forensic and Ancient DNA Studies, *PLoS One*, 10, e0126935.

Hofreiter, M., Paijmans, J.L.A., Goodchild, H., Speller, C.F., Barlow, A., Fortes, G.G., Thomas, J.A., Ludwig, A. & Collins, M.J. 2015. The future of ancient DNA: Technical advances and conceptual shifts, *BioEssays*, 37, 284-93.

Jacobson, P. 2005, 'Investigation: Stalker in the suburbs.', *The Sunday Times*,

Johnston, D. 2006, *The use of DNA in Operation Minstead*, Metropolitan Police Authority.

Kayser, M. & Schneider, P.M. 2009. DNA-based prediction of human externally visible characteristics in forensics: motivations, scientific challenges, and ethical considerations, *Forensic Sci Int Genet*, 3, 154-61.

Kidd, J.R., Friedlaender, F., Pakstis, A.J., Furtado, M., Fang, R., Wang, X., Nievergelt, C.M. & Kidd, K.K. 2011. Single nucleotide polymorphisms and haplotypes in Native American populations, *Am J Phys Anthropol*, 146, 495-502.

Kidd, K.K., Pakstis, A.J., Speed, W.C., Lagacé, R., Chang, J., Wootton, S., Haigh, E. & Kidd, J.R. 2014. Current sequencing technology makes microhaplotypes a powerful new type of genetic marker for forensics, *Forensic Sci Int Genet*, 12, 215-24.

Kidd, K.K., Pakstis, A.J., Speed, W.C., Lagace, R., Chang, J., Wootton, S. & Ihuegbu, N. 2013. Microhaplotype loci are a powerful new type of forensic marker, *Forensic Sci Int Genet Supp Series*, 4, e123-e4.

Kowal, E. & Jenkins, M. 2016, 'DNA Nation raises tough questions for Indigenous Australians', The Conversation, May 25, 2016, viewed 15th February, 2019, <https://theconversation.com/dna-nation-raises-tough-questions-for-indigenous-australians-59877>.

Lipson, M., Szécsényi-Nagy, A., Mallick, S., Pósa, A., Stégmár, B., Keerl, V., Rohland, N., Stewardson, K., Ferry, M., Michel, M., et al. 2017. Parallel palaeogenomic transects reveal complex genetic history of early European farmers, *Nature*, 551, 368.

Ma, K., Zhao, X., Li, H., Cao, Y., Li, W., Ouyang, J., Xie, L. & Liu, W. 2018. Massive parallel sequencing of mitochondrial DNA genomes from mother-child pairs using the ion torrent personal genome machine (PGM), *Forensic Sci Int Genet*, 32, 88-93.

Malyarchuk, B., Derenko, M., Grzybowski, T., Perkova, M., Rogalla, U., Vanecek, T. & Tsybovsky, I. 2010. The peopling of Europe from the mitochondrial haplogroup U5 perspective, *PLoS One*, 5, e10285.

Markham, F. & Biddle, N. 2018. Recent changes to the Indigenous population geography of Australia: evidence from the 2016 Census, *Australian Population Studies*, 2, 1-13.

Martin, B., Blackie, R., Taylor, D. & Linacre, A. 2018. DNA profiles generated from a range of touched sample types, *Forensic Sci Int Genet*, 36, 13-9.

Minister for Justice 2015, 'New National Missing Person and Victim System [Media Release]', <https://crimtrac1-site.govcms.gov.au/sites/g/files/net526/f/documents/150729%20Minister%20for%20Justice%20-%20Media%20release%20-%20New%20National%20Missing%20Perso....pdf?v=1439356138>.

Nielsen, R., Akey, J.M., Jakobsson, M., Pritchard, J.K., Tishkoff, S. & Willerslev, E. 2017. Tracing the peopling of the world through genomics, *Nature*, 541, 302.

Nievergelt, C.M., Maihofer, A.X., Shekhtman, T., Libiger, O., Wang, X., Kidd, K.K. & Kidd, J.R. 2013. Inference of human continental origin and admixture proportions using a highly discriminative ancestry informative 41-SNP panel, *Investig Genet*, 4, 13.

Pakstis, A.J., Fang, R., Furtado, M.R., Kidd, J.R. & Kidd, K.K. 2012. Mini-haplotypes as lineage informative SNPs and ancestry inference SNPs, *Eur J Hum Genet*, 20, 1148-54.

Parson, W. & Bandelt, H.J. 2007. Extended guidelines for mtDNA typing of population data in forensic science, *Forensic Sci Int Genet*, 1, 13-9.

Parson, W., Brandstätter, A., Alonso, A., Brandt, N., Brinkmann, B., Carracedo, A., Corach, D., Froment, O., Furac, I., Grzybowski, T., et al. 2004. The EDNAP mitochondrial DNA population database (EMPOP) collaborative exercises: organisation, results and perspectives, *Forensic Sci Int*, 139, 215-26.

Parson, W. & Dur, A. 2007. EMPOP--a forensic mtDNA database, *Forensic Sci Int Genet*, 1, 88-92.

Phillips, C., Amigo, J., Carracedo, Á. & Lareu, M.V. 2015. Tetra-allelic SNPs: Informative forensic markers compiled from public whole-genome sequence data, *Forensic Sci Int Genet*, 19, 100-6.

Phillips, C., Parson, W., Lundsberg, B., Santos, C., Freire-Aradas, A., Torres, M., Eduardoff, M., Børsting, C., Johansen, P., Fondevila, M., et al. 2014. Building a forensic ancestry panel from the ground up: The EUROFORGEN Global AIM-SNP set, *Forensic Sci Int Genet*, 11, 13-25.

Phillips, C., Prieto, L., Fondevila, M., Salas, A., Gómez-Tato, A., Álvarez-Dios, J., Alonso, A., Blanco-Verea, A., Brión, M., Montesino, M., et al. 2009. Ancestry Analysis in the 11-M Madrid Bomb Attack Investigation, *PLoS One*, 4, e6583.

Quintans, B., Alvarez-Iglesias, V., Salas, A., Phillips, C., Lareu, M.V. & Carracedo, A. 2004. Typing of mitochondrial DNA coding region SNPs of forensic and anthropological interest using SNaPshot minisequencing, *Forensic Sci Int*, 140, 251-7.

Santos, C., Phillips, C., Fondevila, M., Daniel, R., van Oorschot, R.A.H., Burchard, E.G., Schanfield, M.S., Souto, L., Uacyisrael, J., Via, M., et al. 2016a. Pacifiplex: an ancestry-informative SNP panel centred on Australia and the Pacific region, *Forensic Sci Int Genet*, 20, 71-80.

Santos, C., Phillips, C., Gomez-Tato, A., Alvarez-Dios, J., Carracedo, A. & Lareu, M.V. 2016b. Inference of Ancestry in Forensic Analysis II: Analysis of Genetic Data, *Methods Mol Biol*, 1420, 255-85.

Scudder, N., McNevin, D., Kelty, S.F., Walsh, S.J. & Robertson, J. 2018. Forensic DNA phenotyping: Developing a model privacy impact assessment, *Forensic Sci Int Genet*, 34, 222-30.

Shih, S.Y., Bose, N., Gonçalves, A.B.R., Erlich, H.A. & Calloway, C.D. 2018. Applications of Probe Capture Enrichment Next Generation Sequencing for Whole Mitochondrial Genome and 426 Nuclear SNPs for Forensically Challenging Samples, *Genes*, 9, 49.

Spinney, L. 2008. Eyewitness identification: line-ups on trial, *Nature*, 453, 442-4.

Templeton, J.E.L., Brotherton, P.M., Llamas, B., Soubrier, J., Haak, W., Cooper, A. & Austin, J.J. 2013. DNA capture and next-generation sequencing can recover whole mitochondrial genomes from highly degraded samples for human identification, *Investig Genet*, 4, 26.

Turchi, C., Stanciu, F., Paselli, G., Buscemi, L., Parson, W. & Tagliabracci, A. 2016. The mitochondrial DNA makeup of Romanians: A forensic mtDNA control region database and phylogenetic characterization, *Forensic Sci Int Genet*, 24, 136-42.

Walsh, S., Chaitanya, L., Breslin, K., Muralidharan, C., Bronikowska, A., Pospiech, E., Koller, J., Kovatsi, L., Wollstein, A., Branicki, W., et al. 2017. Global skin colour prediction from DNA, *Hum Genet*, 136, 847-63.

Walsh, S., Chaitanya, L., Clarisse, L., Wirken, L., Draus-Barini, J., Kovatsi, L., Maeda, H., Ishikawa, T., Sijen, T., de Knijff, P., et al. 2014. Developmental validation of the HIrisPlex system: DNA-based eye and hair colour prediction for forensic and anthropological usage, *Forensic Sci Int Genet*, 9, 150-61.

Ward, J. 2016, *To investigate specialist DNA techniques for the identification of compromised human remains*, The Winston Churchill Memorial Trust of Australia.

Ward, J. 2018. The past, present and future state of missing persons investigations in Australia, *Aust J Forensic Sci*, 50, 708-22.

Watt, E. & Kowal, E. 2018, 'Why DNA tests for Indigenous heritage mean different things in Australia and the US', The Conversation, October 25, <http://theconversation.com/why-dna-tests-for-indigenous-heritage-mean-different-things-in-australia-and-the-us-105367>.

Wise, R.A., Sartori, G., Magnussen, S. & Safer, M.A. 2014. An Examination of the Causes and Solutions to Eyewitness Error, *Front Psychiatry*, 5, 102.